IRE Transactions

ON AUTOMATIC CONTROL

Volume AC-6 **SEPTEMBER, 1961** **Number 3**

TABLE OF CONTENTS

PUBLISHED BY THE

PROFESSIONAL GROUP ON AUTOMATIC CONTROL

The Issue in Brief

The Sensitivity Problem in Sampled-Data Feedback Systems—I. M. Horowitz

Sensitivity of system performance to parameter variations is often neglected, and yet it is the key to the importance and the use of feedback control. In this paper, sampled-data systems are considered with respect to sensitivity. It is shown how a compromise must be made between the values of the sampling period, the system response, and the sensitivity functions in any design procedure. A numerical design problem with substantial parameter variation is given.

Self-Optimization of a Control System by Means of a Logic Circuit—T. Isobe

A logic circuit is used to provide self-optimization of a control circuit. The device makes successive trials by giving values of a parameter to the system, and, on the basis of the resulting successive observations of the mean square error, it determines the value of the parameter producing the minimum error. The device is also able to follow the change of the system conditions to maintain the optimum.

Stability Conditions of Pulse-Width-Modulated Systems Through the Second Method of Lyapunov—T. T. Kadota and H. C. Bourne, Jr.

Pulsewidth-modulated systems are inherently nonlinear, and, for a rigorous stability study, its dynamic behavior is described by a set of first-order difference equations to which one of the theorems in the second method of Lyapunov may be applied to give a sufficient condition for stability in the large. This is obtained by choosing a positive-definite quadratic form as a Lyapunov function V, with a sufficient condition that the whole space of ΔV is reduced to negative-definiteness. Upon this basis, a systematic procedure of obtaining, analytically, a sufficient condition for asymptotic stability in the large is developed for various types of pulsewidth-modulated systems.

Stability and Graphical Analysis of First-Order Pulse-Width-Modulated Sampled-Data Regulator Systems—E. Polak

This discussion of pulsewidth-modulated sampled-data systems is considerably different from that given in the preceding paper on the same subject. Although it is a simple graphical method of analyzing the stability of a pulse-width-modulated sampled-data system in the large, it is applicable only to very simple systems. However, it may be possible to use it with approximations on higher-order systems to indicate the nature of the system behavior.

Analysis of Pulse-Width Modulated Control Systems—F. R. Delfeld and G. J. Murphy

The two previous papers considered this subject in two different ways. Here, two other methods of analysis are provided. Through the use of the difference equation and the separation of the resulting linear and nonlinear terms, the output at the sampling instants is expressed as a function of the sampled error, and z-transform theory is used to obtain an exact solution for the error at the sampling instants. Then this method of analysis is combined with a modified describing function to investigate stability without neglecting pulse-width saturation. The accuracy and simplicity of the method are illustrated with examples.

Effects of Quantization on Feedback Systems with Stochastic Inputs—R. Kramer

The three previous papers have considered one form of discontinuous signals that may appear in a control system. This paper discusses another form—that of a quantized signal, continuous in time, but not in amplitude. The analysis of the effects of quantization are based on the assumptions that the input signal is Gaussian, and that certain joint error distributions are also Gaussian. It is desired to determine the error autocorrelation as a function of the quantizer for size. A nonlinear integral equation relating the error autocorrelation to the system parameters is developed, an iteration procedure for successive approximations to the solution is presented, several examples are given, and experimental results are shown.

Minimizing Effects of Disturbing Signals Through a Minimum Square-Error Criterion—M. Sobral, Jr.

Besides affording control over parameter sensitivity on system performance as described in the first paper, another important reason for using feedback is in the improvement of disturbing-signal rejection. In one technique, the sum of the command signal plus the disturbing signal, transferred to the input of the system, is used as the input signal. One of the resulting compensating transfer functions formed in the process has to be fixed arbitrarily. Then the optimum over-all transfer function, which minimizes the integral of the square of the error between the desired output and the actual one, is calculated to obtain the remaining compensator. However, the technique does not provide a method for independent determination of the compensators and the required rejection of the disturbing signal may not be obtained. The purpose of this paper is to suggest an analytical technique for determining the compensators with the minimum bandwidth necessary to satisfy a desired over-all transfer function and a required rejection of a disturbing signal. In addition, the technique provides physically realizable compensating transmissions. A brief discussion by O. J. M. Smith illustrates some important points.

Integral Transforms for a Class of Time-Varying Linear Systems—K. S. Narendra

An extension of the transform method to systems having parameters which vary with time is described. If λ is used to represent a general domain, a system function $H(\lambda)$, independent of time, may be defined for the linear system. This function has many of the advantages of the Laplace transform in stationary systems in determining system behavior.

Signal Stabilization of Self-Oscillating Systems—R. Oldenburger and T. Nakada

The self-oscillations of a physical system often may be removed by the introduction of an appropriate stabilizing signal which changes the open-loop gain in a nonlinear manner, and, in general, nonlinear system performance may be improved by the introduction of extra signals. It is shown that, with the aid of Fourier series, the designer can determine the periodic signal to be inserted which will yield a stabilizing input to a nonlinear element in the loop. The approach given explains experimental results.

An Analytical Approach to Root Loci—K. Steiglitz

General algebraic root loci equations are given in polar and Cartesian coordinates, and a synthesis method is suggested which produces linear equations in the coefficients of the open-loop transfer function when closed-loop poles and their corresponding gains are specified. A superposition theorem is presented which shows how the root loci for two open-loop functions place constraints on the locus of their product. With a knowledge of the simple lower-order loci, this theorem is helpful in sketching and constructing root loci.

s-Plane Design of Compensators for Feedback Systems—C. D. Pollak and G. J. Thaler

Like the previous paper, the techniques considered here are based on the root-locus method. The effects of the s-plane poles and zeroes of a compensator on the s-plane gain and phase of an open-loop system are presented as a family of curves. A design technique is developed which permits compensation design to satisfy simultaneous specifications of root location and gain. The method defines the minimum number of compensator sections required, and it gives the logical interpretation of the relative merits of phase lead and lag compensators.

CORRESPONDENCE

This issue of the TRANSACTIONS includes an unusually large number of interesting Correspondence items, ranging from corrections and comments on previous papers to highly technical notes and an interesting adventure in the Soviet Union.

The Sensitivity Problem in Sampled-Data Feedback Systems*

ISAAC M. HOROWITZ†, SENIOR MEMBER, IRE

Summary—This paper is devoted to a discussion of the effect of parameter variations on the system response in sampled-data feedback systems. It is shown that in any single degree of freedom feedback configuration, the system response and especially its overshoot are inherently very sensitive to parameter variation. By means of a suitable transformation, the properties of the sensitivity function can be studied in terms of the usual continuous system frequency concepts, *e.g.*, bandwidth and loop transmission shaping on the Bode plane. There is a basic limitation on the loop transmission bandwidth that can be obtained in any sampled-data feedback configuration. This limitation makes it impossible to secure the unlimited sensitivity reduction which is theoretically available in minimum phase continuous systems. It is shown how one must achieve a compromise between the values of the sampling period, the system response, and the sensitivity function. The design procedure is illustrated in detail with a numerical design problem in which there is substantial parameter variation.

Introduction

THERE is considerable literature concerning the design of sampled-data feedback systems. Most of this literature deals with the problem of specifying a suitable closed-loop system transfer function and its realization under the constraint of a feedback configuration. The problem of the sampled-data system's sensitivity to parameter variations, however, has been almost completely neglected.[1] It is well known that one of the principal reasons for using feedback is to reduce the sensitivity of the system response to parameter variations. It is surely important, therefore, that sensitivity considerations and requirements should appear as an integral quantitative part of the design procedure. It has been shown[2,3] that in both the single variable and the linear, multivariable minimum phase feedback control system, it is possible to reduce to any desired extent the system's sensitivity to parameter variations, no matter how large the latter may be. Straightforward quantitative design procedures for achieving this reduction have been presented. The important basic question is whether unlimited sensitivity reduction is also possible in sampled-data systems, or whether the sampling introduces some limitation.

* Received by the PGAC, October 25, 1960; revised manuscript received, March 3, 1961.

† Hughes Research Labs., A Div. of Hughes Aircraft Co., Malibu, Calif.

[1] S. F. Schmidt, "Application of continuous system design concepts to the design of sampled-data systems," *Trans. AIEE*, vol. 78 (*Applications and Industry*, no. 42), pp. 74–79; May, 1959.

[2] I. M. Horowitz, "Fundamental theory of automatic linear feedback control systems," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-4, pp. 5–19; December, 1959.

[3] I. M. Horowitz, "Synthesis of linear, multivariable feedback control systems," *Proc. Natl. Electronics Conf.*, vol. 15, pp. 276–289; 1959.

Sensitivity of the Single Degree of Freedom Structure

The structure illustrated by Fig. 1 consists of the plant, or controlled process P, and the compensation network, or digital controller G. In the typical control problem, the plant is given, which leaves the choice of G as the only available freedom. It is obvious, therefore, that only one system function may be obtained independently. In the literature on sampled-data systems, G is always chosen so that the desired system transfer function is obtained. The designer must then accept the resulting sensitivity function. Thus, in Fig. 1, let C_0, T_0, P_0 be the output, transfer, and plant functions, respectively, when the plant has its nominal design value, and let C, T, P be the corresponding functions at a new plant value P. Also let

$$\Delta P = P - P_0 \tag{1a}$$

$$\Delta T = T - T_0 \tag{1b}$$

$$\Delta C = C - C_0. \tag{1c}$$

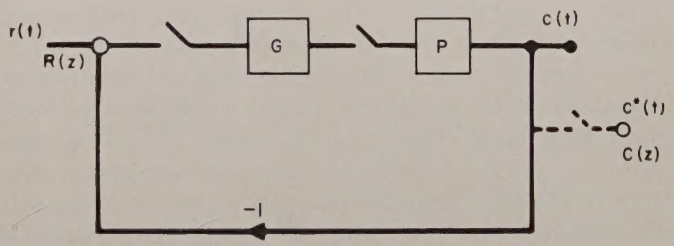


Fig. 1—Structure with single degree of freedom.

Since

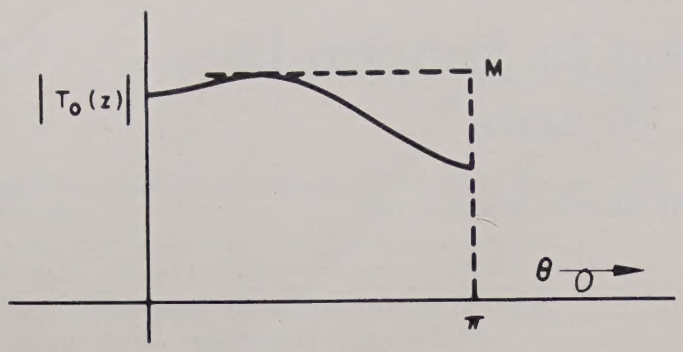
$$T_0(z) = \frac{G(z)P_0(z)}{1 + G(z)P_0(z)}, \tag{2}$$

it is easily found that[4]

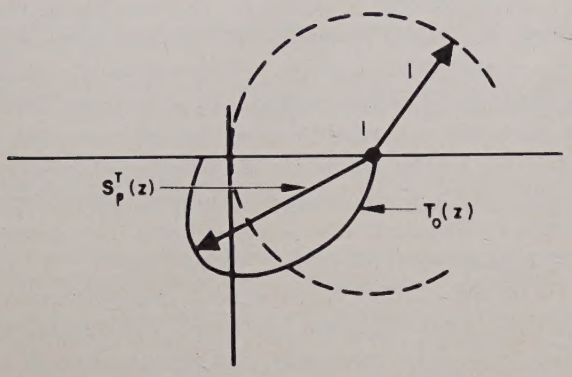
$$S_P^T \triangleq \frac{\Delta T/T}{\Delta P/P} = \frac{1}{1 + G(z)P_0(z)} = 1 - T_0(z). \tag{3}$$

Consider the locus of S_P^T as $z = e^{j\theta}$ varies over the upper half of the unit circle in the z-plane. A typical $|T_0(z)|$ is sketched in Fig. 2(a), and its polar locus is shown in Fig. 2(b). The polar locus of the sensitivity function in Fig. 2(b) is easily obtained from (3). It is the vector originating from the point (1, 0) and ter-

[4] The symbol $\triangleq$ means "equal by definition."



(a)



(b)

Fig. 2—(a) A typical $|T_0(z)|$. (b) Relation between $T(z)$ and $S_P{}^T(z)$ in the single degree of freedom structure.

minating on $T_0(z)$. Over an important range of z (important because $T_0(z)$ has significant magnitude in this range), the sensitivity of the system is greater than one. This means that over an important part of the transform variable range, the system is more sensitive to plant parameter variations than it would be if no feedback were used. This range in which $|S|>1$ is associated with the system step response overshoot. The greater the step response overshoot, and consequently the value of M in Fig. 2(a), the greater is the value of the sensitivity function in this range of z.

It is not difficult to obtain the effect of plant parameter variations on the output pulse series. Let

$$C_0(z) = \sum_{n=1}^{\infty} c_n z^{-n} \tag{4a}$$

$$T_0(z) = \sum_{1}^{\infty} t_n z^{-n} \tag{4b}$$

$$C(z) = \sum_{1}^{\infty} c_n' z^{-n} \tag{4c}$$

and

$$\Delta C(z) = \sum_{1}^{\infty} \tau_n z^{-n} \tag{5}$$

with

$$\tau_n = c_n' - c_n. \tag{6}$$

If (3) is expanded in a power series in z^{-1} and if the corresponding terms are equated, the result (for the special case $\Delta P/P = K$ is independent of z) is

$$\tau_1 = \frac{K}{1-K}, \qquad \tau_2 = \frac{K}{1-K}(c_2 - t_1 c_1') \tag{7a}$$

and in general

$$\tau_n = \frac{K}{1-K}(c_n - c_{n-1}' t_1 - c_{n-2}' t_2 \cdots - c_1' t_{n-1}). \tag{7b}$$

More generally, when

$$\frac{\Delta P}{P} = \sum_{0}^{\infty} \delta_n z^{-n}, \tag{8}$$

then

$$\left.\begin{aligned} &\tau_1 = \frac{\delta_0}{1-\delta_0} c_1, \qquad \tau_2 = \frac{1}{1-\delta_0}(\delta_0 c_2 + \delta_1 c_1' - \delta_0 c_1' t_1), \\ &\text{and in general} \\ &\tau_n(1-\delta_0) = \delta_0 c_n + \sum_{1}^{n-1} \delta_i c_{n-i}' - t_1 \sum_{0}^{n-2} \delta_i c_{n-i-1}' \\ &\qquad - t_2 \sum_{0}^{n-3} \delta_i c_{n-i-2}' \cdots - t_{n-1}\delta_0 c_1'. \end{aligned}\right\} \tag{9}$$

It is easier to obtain the series for $C_0/\Delta C$. Straightforward manipulation of (3) leads to

$$\frac{C_0}{\Delta C} = 1 - \frac{\Delta P}{P}(1 - T_0). \tag{10}$$

The work involved in calculating the right side of (10), its inversion, and final multiplication by C_0 is usually less than that involved in evaluating (9). This is illustrated with an example[5] in which

$$T_0(z) = \frac{(z + 0.45)}{(z^2 + 0.25z + 0.2)}.$$

The nominal step-function response sketched in Fig. 3 is

$$C_0(z) = z^{-1} + 1.2z^{-2} + 0.95z^{-3} + 0.973z^{-4} + 1.017z^{-5} + 1.001z^{-6} + \cdots,$$

and

$$T_0(z) = z^{-1} + 0.2z^{-2} - 0.25z^{-3} + 0.023z^{-4} + 0.44z^{-5} - 0.016z^{-6} + \cdots.$$

Suppose the plant gain constant increases by 25 per cent so that $\Delta P/P = K = 0.20$. Then from (7),

$$\Delta C(z) = 0.25z^{-1} - 0.0125z^{-2} - 0.122z^{-3} + 0.055z^{-4} + 0.023z^{-5} - 0.076z^{-6} + \cdots.$$

[5] This value of $T_0(z)$ is taken from J. G. Truxal, "Control System Synthesis," McGraw-Hill Book Co., Inc., New York, N. Y., p. 540; 1955.

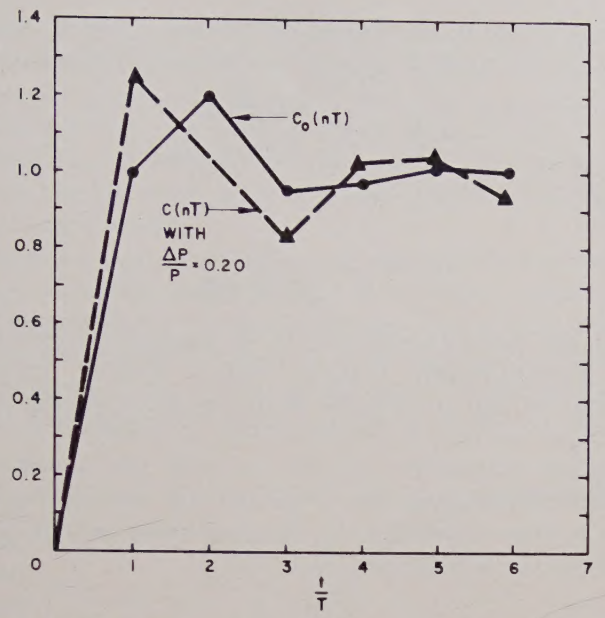


Fig. 3—Comparison of nominal and perturbed step response.

The resulting $C(z) = C_0(z) + \Delta C(z)$ is also sketched in Fig. 3. It is seen that the system step response is quite sensitive to parameter changes.

A system whose output is very sensitive for one type of input function is not necessarily sensitive for all input functions. In fact, almost any feedback control system output is highly sensitive to one class of input function and highly insensitive to some other class of input function. This is clarified by writing (3) in the form

$$\Delta C = \frac{\Delta P}{P}(1 - T_0). \tag{11}$$

If C is comparatively large in the region where $1 - T_0$ is small and comparatively small in the region where $1 - T_0$ is large, then ΔC is small. Thus, even though $\Delta C/C$ is independent of C, the region in which $\Delta C/C$ is large can be such that its effect on the time response is small.

The above remarks are clarified with the aid of the following example:[6]

$$T(z) = 0.73z^{-1} + 1.35z^{-2} - 0.90z^{-3} - 0.18z^{-4}.$$

This is the system transfer function of a minimal prototype response system for a plant transfer function $P(z)$, which has a double zero outside the unit circle at $z = 2.34$. The system responds to a step and ramp input with no ripple in the steady state, and the transient has the shortest possible finite duration. The step and ramp responses for this system are sketched in Figs. 4(a) and 4(b), respectively. Now, assume a 10 per cent increase in the plant gain constant. When (9) is used, the effect on the response is found for both cases; this effect is plotted in Figs. 4(a) and 4(b). Clearly, the

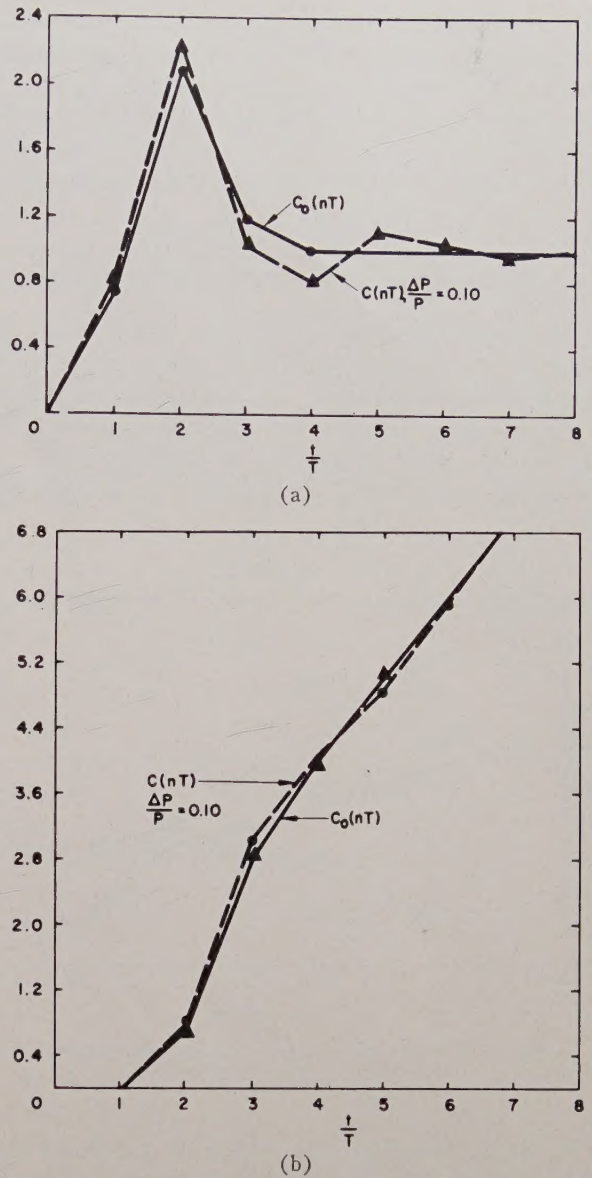


Fig. 4—(a) Minimal prototype system—comparison of nominal and perturbed step response. (b) Minimal prototype system—comparison of nominal and perturbed ramp response.

step response sensitivity is relatively greater than that of the ramp response. The reason for this difference in sensitivity is that the ramp response is relatively stronger than the step response in the "frequency" range in which the sensitivity function is small. In making this comparison, it is convenient to use a new variable[7]

$$w = u + jv \triangleq \frac{z - 1}{z + 1} = \frac{e^{sT} - 1}{e^{sT} + 1}. \tag{12}$$

The $j\omega$-axis in the s-plane and the unit circle in the z-plane map onto the jv-axis in the w-plane. The left half of the s-plane maps on the left half of the w-plane.

[6] J. F. Ragazzini and G. F. Franklin, "Sampled-Data Control Systems," McGraw-Hill Book Co., Inc., New York, N. Y., p. 170; 1958.

[7] C. W. Johnson, D. P. Nordling, and D. P. Lindorff, "Extension of continuous-data system design techniques to sampled data control systems," *Trans. AIEE*, vol. 74 (*Application and Industry*, no. 20), pp. 252–263; September, 1955.

The transcendental pulsed transfer function in s, which becomes a rational function in z, is also a rational function in w, but there is the advantage that the exceedingly elegant and powerful Bode plot and loop-shaping techniques can be used. The jv-axis can be considered as a generalized frequency axis. In the present example,

$$S(w)=\frac{1.25w(w+0.0025)(w+0.8)}{(w+0.0633)[w^2+(2)(0.854)(0.418)w+(0.418)^2]},$$

and its magnitude (asymptotic) for $w=jv$ is sketched in Fig. 5. The w transforms of the step and ramp outputs are also sketched (they are normalized to the same infinite frequency value for a fair comparison). It is clear that the ramp transform is relatively larger over that frequency region in which the sensitivity is smallest.

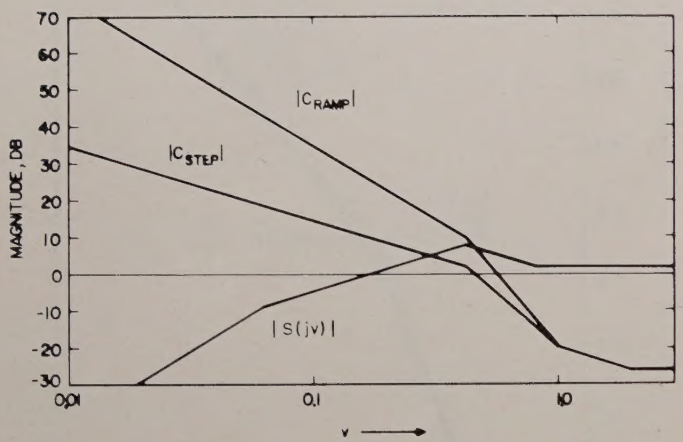


Fig. 5—Comparison of sensitivity, step output, and ramp output frequency responses.

The above technique for obtaining a desired insensitivity with the single degree of freedom configuration of Fig. 1 is in most cases impractical, because the system transfer function bandwidth (in v) is made greater than is actually necessary. In most cases, the system bandwidth is made as small as possible in order to discriminate against the noise that enters with the useful signal.

What then can be done to secure any desired insensitivity simultaneously with a desired $T_0(z)$? In the continuous system, any two-degree-of-freedom configuration can be used, and design procedures for this purpose have been developed.[2,3] Can these procedures be readily extended to sampled-data systems, or is there any inherent limitation in the sensitivity reduction achievable?

Sensitivity Limitations in the Two-Degree-of-Freedom Structures

Consider any two-degree-of-freedom configuration in which the feedback signal is sampled; *e.g.*, see Fig. 6. There are now two degrees of freedom G and H, which apparently make it possible to realize independently

$$T(z)=\frac{G(z)P(z)}{1+G(z)P(z)H(z)} \tag{13}$$

and

$$S_P{}^T=\frac{1}{1+G(z)P(z)H(z)}. \tag{14}$$

Can any desired sensitivity of the sampled output then be simultaneously realized with a desired nominal $T_0(z)$? In the continuous feedback problem, the sensitivity function is expressed as a function of the Laplace transform variable s, and the sensitivity specifications along the $s=j\omega$ axis are used to control the system sensitivity. In the continuous system with $F(s)$ as the transform of $f(t)$

$$f(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}F(\omega)e^{j\omega t}d\omega. \tag{15}$$

The argument is simply that if the changes in $F(\omega)$ over its important range are small, then the corresponding changes in $f(t)$ must also be small. This argument is the basis for using frequency response methods in the design of feedback amplifiers and feedback control systems. No one as yet has shown how to control directly the time response sensitivity.

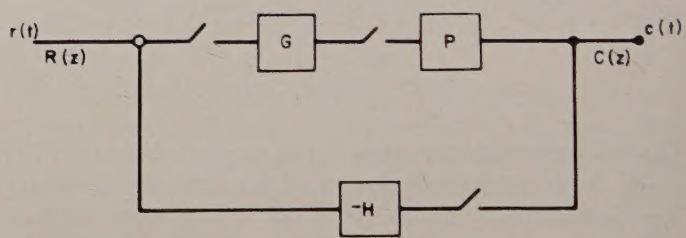


Fig. 6—Structure with two degrees of freedom.

The analogous variable in sampled-data systems would appear to be $z=e^{sT}$, with the unit circle $z=e^{j\theta}$ corresponding to the $s=j\omega$ axis. It is again more convenient, however, to work with the variable defined by (12). The principal reason for the usefulness of the variable w is that techniques similar to the Bode plot method become available. Furthermore, it will be seen that the inherent limitations on the sensitivity caused by the sampling are unusually well defined when the functions are expressed in terms of $w=jv$. The expression corresponding to (15) is

$$f(nT)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{(1+jv)^{n-1}}{(1-jv)^{n+1}}F(v)dv. \tag{16}$$

The argument used in (15) also applies here, *i.e.*, if $F(v)$ does not vary much in its significant range, the corresponding variation in $f(nT)$ will also be small.

There is in (16) the weighting factor $(1-jv)^{-2}$, which ensures convergence of the integral. If $F(v)=T(v)$, then large v is associated with small $T(v)$. Consequently, the frequency response (with v taking the place of ω) approach to controlling the system time response sensitivity applies here with as much justification as in continuous systems.

Are there any limitations on the loop transmission $L(w)=G(w)H(w)P(w)$ of Fig. 7, or can its poles and zeros be selected freely? It is known that $L(z)$ must have more poles than zeros in order that $L(s)$ go to zero at infinity. If there is to be no pure time lag in $L(s)$, $L(z)$ should have an excess of only one pole over zeros.

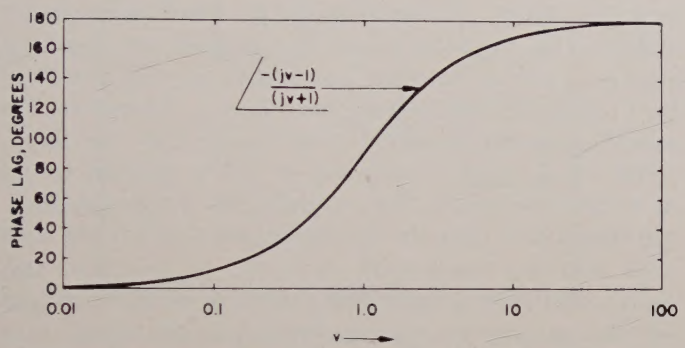


Fig. 7—Phase lag due to $-(jv-1)/(jv+1)$.

Now, when (12) is used, a factor $(w+a)$ in $L(w)$ becomes $[z(1+a)-(1-a)]/z+1$ in $L(z)$.[8] To ensure that $L(z)$ has only one more pole than zeros, $L(w)$ must have in its numerator the factor $-(w-1)$. This factor results in a positive 6-decibel per octave slope in $|L(w)|$ accompanied by the phase lag normally associated with a pole. It, therefore, has a strongde stabilizing effect. Suppose a pole at -1 is assigned to $L(w)$ to cancel the effect of the $-(w-1)$ factor on $|L(w)|$. This combination of pole and zero contributes nothing to the amplitude of $L(w)$, but it does contribute appreciable phase lag, as shown in Fig. 7. The severe restrictions this phase lag imposes on the permitted gain-bandwidth product of $L(w)$ can be seen immediately. It appears impossible to extend the crossover frequency (at which the amplitude crosses the zero-decibel line) much beyond $v=1$. The exact maximum gain-bandwidth restrictions for specified stability margins can be worked out using techniques contained in Bode.[9] $L(w)$ has as many zeros as poles, unless $L(z)$ is assigned a zero at $z=-1$. This is obvious by noting that $z=-1$ corresponds to infinite w. In order to secure a maximum gain-bandwidth product for $L(w)$, it is better to let $L(w)$ have as many zeros as poles. In such a case, the phase lag of $L(v)$ approaches 180° as v approaches infinity. Otherwise, if $L(z)$ has a zero at $z=-1$, the phase lag approaches 270°.

Suppose all the component parts of $L(s)$, *i.e.*, $G(s)$, $H(s)$, and $P(s)$ in Fig. 6, are to consist of the usual continuous-type elements as opposed to pulsed circuits or digital controllers.[10] In realizing the corresponding network, it is then advantageous to expand each of the corresponding pulsed transfer functions in a series of the form $\sum A_i z/(z-a_i)$ and to obtain from standard tables the corresponding continuous transfer functions. For the configuration of Fig. 6 then, $L(z)$ should have a zero of order three at the origin, and, consequently, $L(w)$ must have this order zero at -1. On the other hand, if the plant is to be preceded by a hold circuit, a zero of order two is sufficient. If, in addition, the other building blocks are to consist of pulsed or digital circuits, or will be preceded by hold circuits, no zeros at $w=-1$ are needed.

If it is desired that the poles of $L(s)$ be confined to the negative real axis, then the poles of $L(w)$ are restricted to lie on the negative real w-axis between the origin and -1. In the configuration illustrated in Fig. 6, there is no need to avoid complex highly underdamped poles of $L(w)$, because they can be assigned to $G(s)$ or $H(s)$ and therefore will not adversely affect intersample behavior. In the structure of Fig. 8, however, they do affect the intersample behavior. The relation between the w- and s- planes is shown in Fig. 9, where loci of constant $\alpha=\sigma/\omega_s$ and constant $\beta=\omega/\omega_s$ are drawn; the customary notation is used, *viz.*, $s=\sigma+j\omega$, $\omega_s=2\pi/T$, and T is the sampling period. A few loci of constant damping factor are also sketched. If the structure shown in Fig. 8 is used, it should be noted that poles and zeros of $L(s)$ determine the intersample behavior, which means, generally, that $L(w)$ should have no poles beyond about -1.2.

Because of the compulsory $-(w-1)$ factor, there is a limit on the gain-bandwidth product of $L(w)$ that can be achieved. Let us explore the limits under various constraints. Suppose $L(w)$, excluding the $-(w-1)/(w+1)$ part, consists only of one pole at the origin. If a 30° phase margin is desired (see Fig. 7), then the crossover "frequency" is $v=0.58$; for 40°, it is 0.35, etc. This bandwidth can be extended by letting $L(w)$ have as many zeros as poles. The effect of this additional zero depends on the gain margin desired. It is easier to work backward here; *e.g.*, the crossover "frequency" of $v=1$ associated with a phase margin of 30° determines a specific maximum gain margin, which is found as follows. The pole at the origin and the $-(w-1)/(w+1)$ combination result in a 180° phase lag at $v=1$. For a 30° phase margin, the zero of $L(w)$ must be at $v=1.73$,

[8] Note here that $L(w)$ is not $L(z)|_{z=w}$, just as $L(z)$ is not $L(s)_{s=z}$. The standard notation is used here, *viz.*,

$$L(z)=L^*(s)|_{z=e^{sT}}$$

and

$$L(w)=L(z)|_{z=1+w/1-w}.$$

[9] H. W. Bode, "Network Analysis and Feedback Amplifier Design," D. Van Nostrand Co., Inc., New York, N. Y., ch. 14, pp. 303–336; 1945.

[10] Ragazzini and Franklin, *op. cit.*, pp. 136, 145.

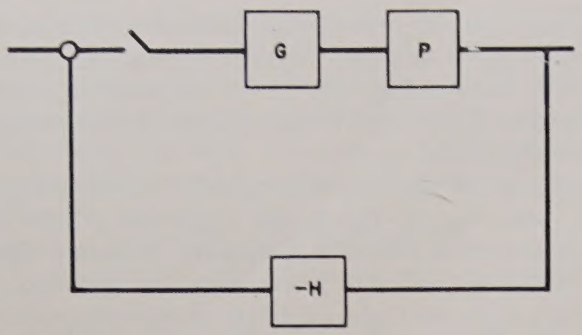


Fig. 8—Configuration in which highly underdamped poles should be avoided.

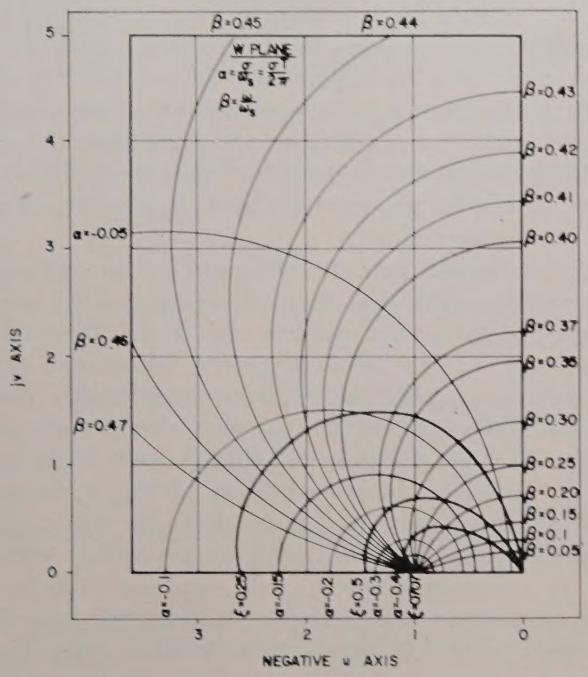


Fig. 9—Mapping of $s=\sigma+j\omega$ plane into $w=u+jv$ plane.

and the resulting gain margin and infinite frequency attenuation is 4.9 db. For the same phase margin, with crossover at $v=0.8$, the gain margin is 6.6 db and the infinite frequency attenuation is 9.6 db.

Suppose an $L(w)$ with one additional pole-zero pair is considered. One zero may be assigned at such a low frequency that the effect on the phase of the pole at the origin is almost completely canceled. The real problem is then: What gain-bandwidth product is achievable with one pole-zero pair for a given phase margin? Because of the properties shown in Fig. 7, the answer is a function of the crossover frequency. For example, a 20-db gain level with crossover at 1.0 (half-power point approximately at 0.1) requires that the final zero be at $v=2.2$ for 30° phase margin with a gain margin and infinite frequency attenuation of 6.8 db. These figures give an indication of the levels of loop gain and crossover frequency that can be achieved.

In view of the severe constraint on the loop gain bandwidth that is achievable, it appears that independent realization of a desired system function $T(w)$ and a desired $S_P{}^T(w)$ is not truly possible. In fact, in order to make $S_P{}^T(w)$ small over the entire significant range of $T(w)$, it is necessary that the "bandwidth" of $T(w)$ be several octaves below $v=1$. In effect, this means that if a small $S_P{}^T(jv)$ is stipulated over the significant range of $T(jv)$, the latter must be severely restricted. Nevertheless, the joint achievement of a desired system time response and its insensitivity to plant parameter variation is still possible.

Suppose a system step response based on that shown in Fig. 10 is desired. In a sampled-data system, this objective could be realized, presumably, over some definite range of sampling period; thus either τ_1 or $\tau_2=2\tau_1$ could be used. If τ_2 is used, the resulting system transfer function $T_2(w)$ will be required to have some bandwidth v_2. On the other hand, if τ_1 is used, the resulting system transfer function bandwidth v_1 will certainly turn out to be considerably less than v_2. Therefore, the sampling period could be chosen such that the bandwidth of the system response in the w domain would be sufficiently small. For example, the simple exponential response e^{-t} has the transform $w+1/w+0.632$ with $\tau=1$ and $w+1/w+0.245$ at $\tau=\frac{1}{2}$. In practice, this means that after a maximum sampling period sufficient for the purpose of reproducing the input signal to a desired extent has been picked, the sampling period must be made somewhat smaller in order to make the system response fairly insensitive to parameter variation. This sacrifice, either in sensitivity or in the maximum permissible sampling period, appears to be one of the fundamental limitations imposed by sampling.

The sacrifice in the sensitivity or in the sampling period may be avoided by using the usual continuous feedback to whatever extent possible. Thus, as far as sensitivity reduction is concerned, the structure in Fig. 11 (assuming, of course, that it is possible to use it) is superior to that in Fig. 8.

The design procedure for simultaneously realizing a desired system response and a desired response insensitivity to plant parameter variations will now be presented and illustrated by means of a design example.

Design for Simultaneous Control of $T(w)$ and $S_P{}^T(w)$

The design procedure is presented by means of a specific numerical problem. The plant with a zero-order hold has the transfer function $aA(1-e^{-sT})/s^2(s+a)$, with A and a varying independently over the range 1 to 2 and 0.5 to 1, respectively. The configuration shown in Fig. 6 is to be used. The bandwidth of $T(w)$ is to extend approximately to $v=0.3$. It is believed that sufficient insensitivity will be attained if, despite the parameter variations, $|T(v)|$ does not vary more than zero at $v=0$, 10 per cent at 0.1, and about 20 per cent at 0.3. At higher frequencies, there is some concern about the maximum peaking in $T(v)$; *e.g.*, $|T(v)|$ shown in Fig. 12 is undesirable. Any such peaking due to parameter variation should be kept to a few decibels.

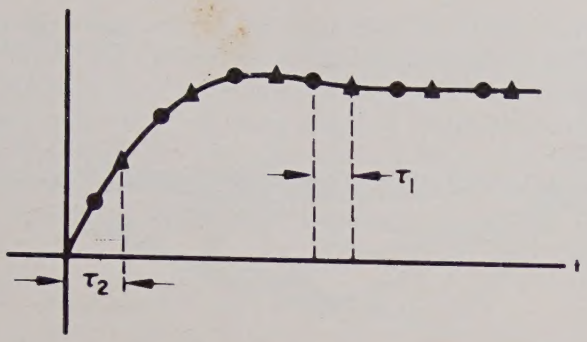


Fig. 10—Illustrative system step response.

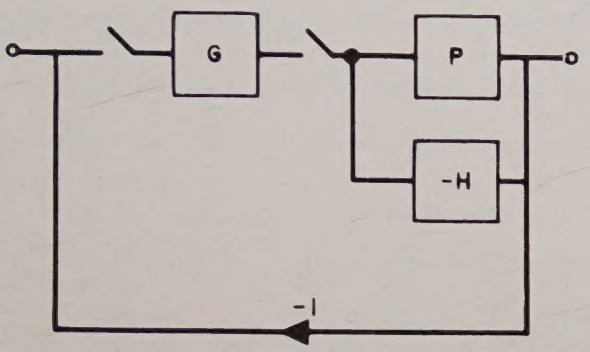


Fig. 11—A structure that is superior in sensitivity-reduction potential.

The sensitivity problem will be examined first. With the use of (13), it can be seen that

$$\frac{T_0(jv)}{T(jv)} = \frac{\frac{P_0}{P}(jv) + L_0(jv)}{1 + L_0(jv)}, \tag{17}$$

where

$$L_0(w) = G(w)P_0(w)H(w) \tag{18a}$$

$$L(w) = G(w)P(w)H(w). \tag{18b}$$

Eq. (17) is very convenient for picking the loop transmission $L_0(w)$ to satisfy the sensitivity specifications. Suppose, for example, that at $v=v_1$, the entire range of variations of $P_0(jv_1)/P(jv_1)$ is contained within the shaded area indicated in Fig. 13 and that at v_1, $L_0(jv_1)$ has the value indicated by the point A in Fig. 13. Then, from (17),

$$\frac{T_0}{T}\frac{(jv_1)}{(jv_1)} = \frac{AB}{AC},$$

and the extreme values of $T_0/T(jv_1)$ are easily obtained.

Conversely, given the complete area of variation of $P_0/P(jv_1)$ and the extreme permissible values of $|T_0/T(jv_1)|$, the range of permissible $L_0(jv_1)$ can be found easily. This is illustrated in Fig. 14, where an assumed boundary of $P_0/P(jv_1)$ has been drawn. Suppose it is desired that $0.80 < |T_0/T(jv_1)| < 1.20$. The cross-hatched region indicates the permissible location of $-L_0(jv_1)$ in the upper half plane. Similar boundaries of permissible locations of $L_0(jv)$ may be sketched at

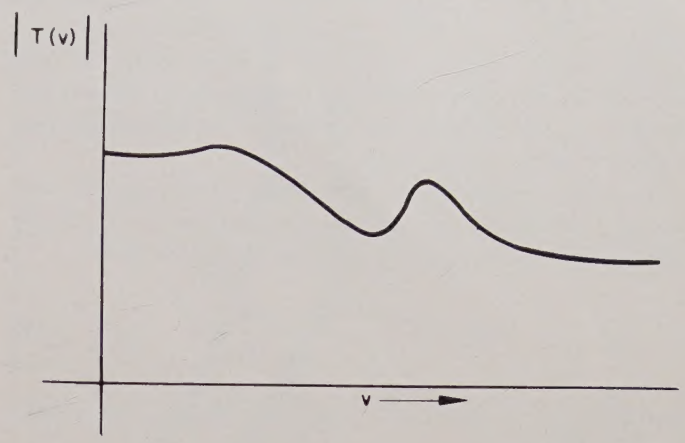


Fig. 12—An unsatisfactory perturbed system response.

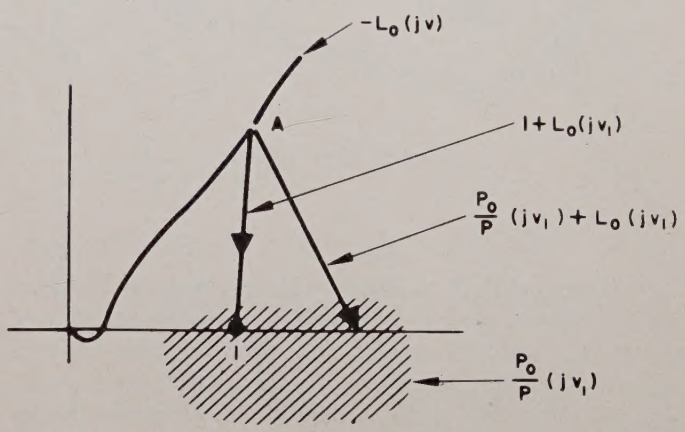


Fig. 13—Construction for determining range of $T_0/T(jv)$.

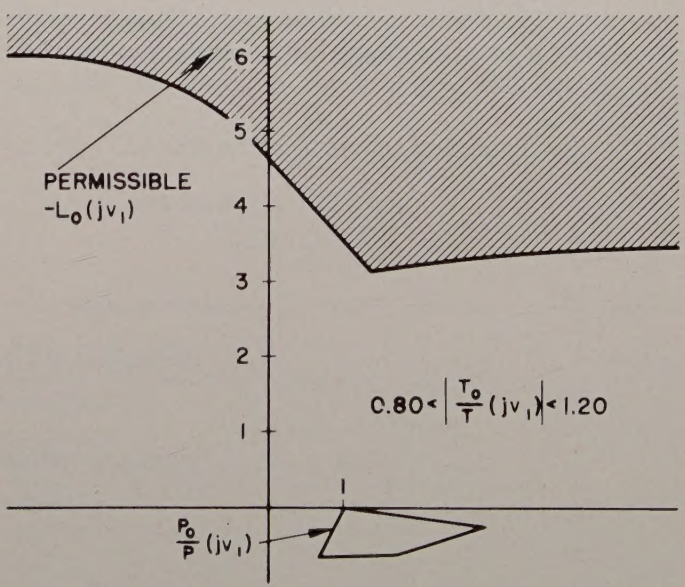


Fig. 14—Determination of permissible range of $L_0(jv_1)$ to maintain specified tolerances on $|T_0/T(jv_1)|$.

other values of "frequency," the resulting boundaries depending on the range of $P_0/P(jv)$ and the specified limits on $T_0/T(jv)$. With these boundaries, it is a relatively simple matter to obtain a satisfactory $L_0(jv)$, providing, of course, that the requirements do not exceed the basic capabilities, in view of the previously discussed limitations on $L_0(jv)$.

In applying the above procedure to the specific numerical problem, the region of variation of $P_0/P(jv)$ for a number of values of v is first obtained. Here,

$$P(s) = \frac{1 - e^{-sT}}{s} \frac{Aa}{s(s + a)}, \tag{19}$$

and if $a_0=0.5$, $A_0=1$, $T=2$ sec are chosen, the resulting $P_0/P(w)$ is

$$\frac{P_0}{P}(w) = \frac{\left(w + \frac{1 - e^{-2a}}{1 + e^{-2a}}\right)(-0.078w^2 - 0.389w + 0.463)}{A(w + 0.463)\left\{ w^2\left[\frac{2}{a(1 + e^{-2a})} - \frac{1}{a} - 1\right] + w\left[\frac{e^{-2a}\left(2 + \frac{1}{a}\right) - \frac{1}{a}}{1 + e^{-2a}}\right] + \left[\frac{1 - e^{-2a}}{1 + e^{-2a}}\right]\right\}}. \tag{20}$$

At a fixed value of A and $w=jv$, the locus of $P_0/P(w)$ as a function of a is obtained by simply calculating P_0/P for a few values of a. The effect on this locus of varying A is obvious, since A appears only as a multiplier in (20). This in Fig. 15, E, F, G, H marks the region of variation of $P_0/P(w)$ for $w=j3.5$. The calculations and sketches are repeated for several values of jv, and the results are shown in Fig. 15. EHA is the locus of $P_0/P(jv)\big|_{v\to\infty}$. Next, the boundaries of the permissible values of $L_0(jv)$ are obtained (in the manner of Fig. 14); see Fig. 16. The requirements on $L_0(w)$ are then apparent.

A simple $L_0(w)$ that is used as a first trial is

$$L_0(w) = \frac{-0.0625(w - 1)(w + 4)}{w(w + 1)}.$$

Its Bode sketch is shown in Fig. 17, and polar sketches at low and high frequencies are shown in Figs. 16 and 18, respectively. While this $L_0(w)$ is more than satisfactory around $v=0.10$, it is quite poor at higher frequencies (30 per cent maximum increase at $v=0.35$, 90 per cent at $v=0.50$). In order to obtain a $L_0(w)$ with somewhat more gain at $v\sim0.3$, some sacrifice in the lower frequency gain is made with

$$L_0(w) = \frac{-0.063(w + 0.1)(w - 1)(w + 4)}{w(w + 0.4)^2}.$$

This too is sketched in Figs. 16–18. At $v=0.35$, the variation in T is 9 per cent, which is higher than that for the first trial, but its performance at $v\sim0.4$ (20 per cent maximum at $v=0.35$, 35 per cent at $v=0.5$) is satisfactory and it is therefore used in the design.

To complete the design, the system transfer $T_0(w)$ is needed. The problem of the choice of the system transfer function in sampled-data systems is difficult and has been the subject matter of many papers. This paper is not concerned with the above problem but rather with that of obtaining the desired insensitivity of the chosen function to plant parameter variations. In this specific example, it is assumed that the function in its significant frequency region is given by $1/(w^2+0.45w+0.10)$. Therefore,

$$T_0(w) = \frac{-K(w - 1)M(w)}{(w^2 + 0.45w + 0.10)},$$

with $M(w)$ to be chosen later. For the configuration of Fig. 6,

$$L_0(w) = GP_0H = \frac{-0.063(w + 0.1)(w - 1)(w + 4)}{w(w + 0.4)^2},$$

$$G = T_0\frac{(1 + L_0)}{P_0} = \frac{12.5K(w + 0.071)(w + 0.463)(w^2 + 0.574w + 0.379)M(w)}{(w^2 + 0.45w + 0.10)(w + 0.4)^2(w + 6.0)},$$

and

$$H = \frac{L_0}{GP_0} = \frac{L_0}{T_0(1 + L_0)} = \frac{0.067}{K}\frac{(w + 0.1)(w^2 + 0.45w + 0.10)(w + 4)}{M(w)(w + 0.071)(w^2 + 0.574w + 0.379)}.$$

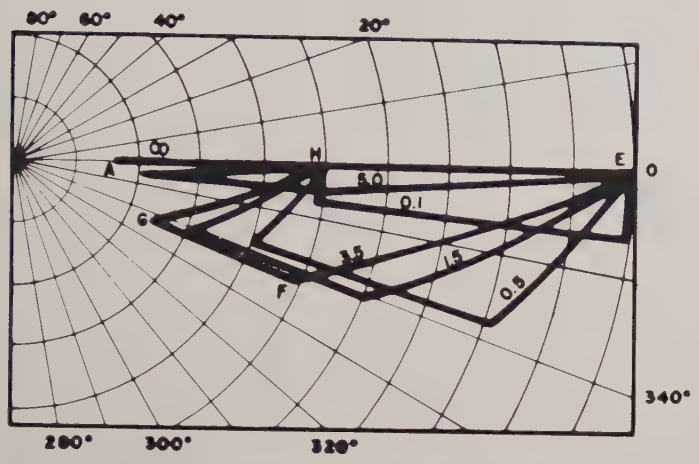


Fig. 15—Range of $P_0/P(jv)$ for several values of v.

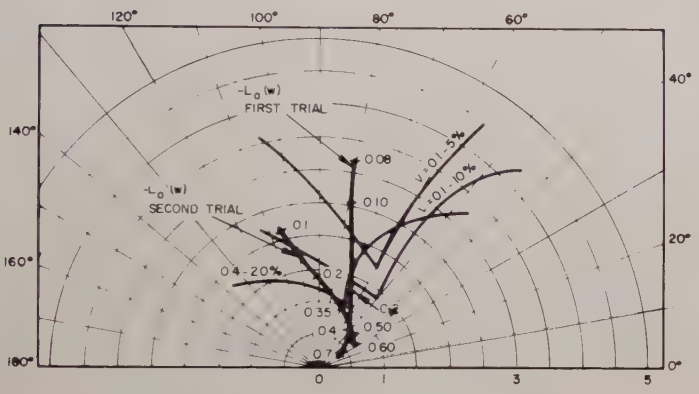


Fig. 16—Boundary of permissible $L_0(jv)$.

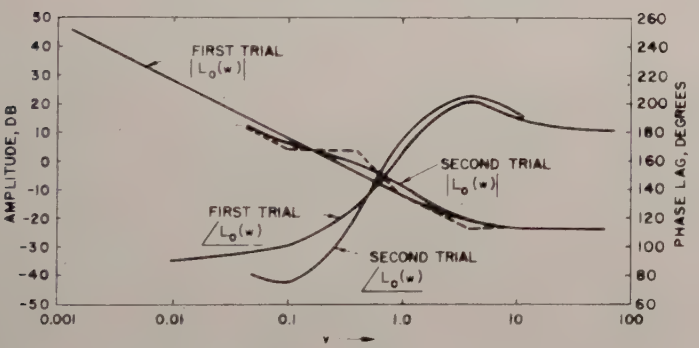


Fig. 17—Bode sketches of $L_0(jv)$.

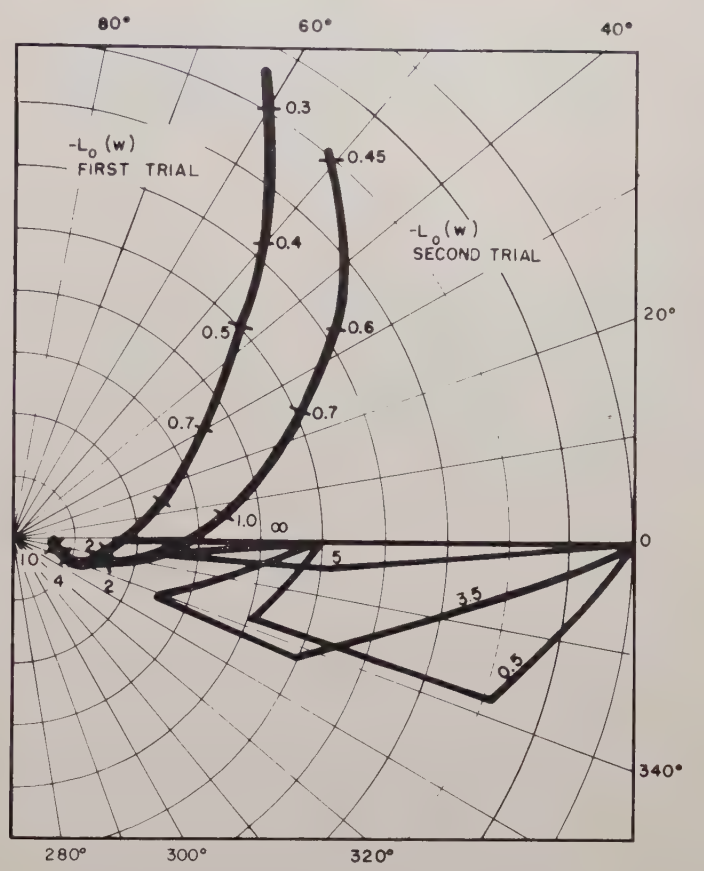


Fig. 18—Determination of $T_0/T(jv)$ at high frequencies.

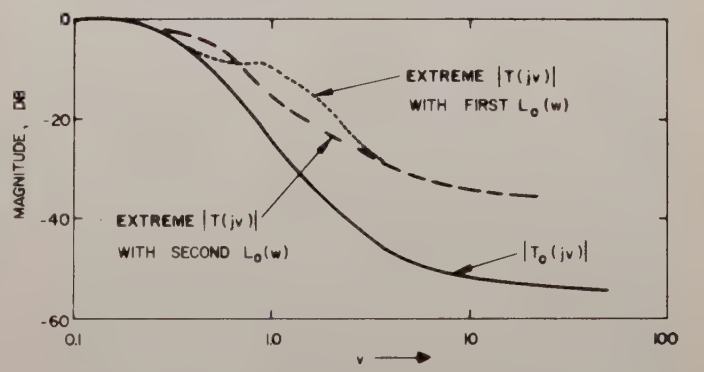


Fig. 19—Extreme value of $|T(jv)|$ due to parameter variation.

Now $M(w)$ is selected so that the maximum peaking of $T(w)$ at higher frequencies is kept within the desired bounds. Thus at $v=1.5$, it is found from Fig. 18 that there is a maximum increase of 18 db in $|T(j1.5)|$. If $|T(j1.5)|$ is never to be more than -15 db, an additional 10-db attenuation is needed. From such considerations, $M(w)$ is selected to be $(w+4)(w+6)/(w+0.463)$, and $K=0.0463/24$ in order that $T_0(0)=1$. The resulting extreme peak values of $T(jv)$ for the two trial values of $L(jv)$ are shown in Fig. 19.

The balance of the design procedure is straightforward. Since realization techniques for sampled-data transfer functions are available in z-transform notation, $G(w)$ and $H(w)$ are converted into the z notation and the resulting transfer functions are obtained.[10]

Conclusions

A feedback system in which there is a sampler in the loop and which has a minimum of two lags, is inherently a nonminimum phase system. This property restricts the amount of loop gain-bandwidth that is realizable and consequently limits the amount of sensitivity reduction to parameter variation.

The transformation $w=(z-1)/(z+1)=(e^{sT}-1)/(e^{sT}+1)$, permits the powerful and elegant Bode techniques and the fundamental Bode feedback theory, to be applied to sampled-data feedback systems. Furthermore, the detailed quantitative design procedure for limiting the effect of plant parameter variation, which has been developed for continuous systems, is thereby made available for sampled-data feedback systems.

Self-Optimization of a Control System by Means of a Logic Circuit*

TAKASHI ISOBE†

Summary—**The self-optimization of a control system has been tried by means of a logic circuit. The device makes successive trials by giving values of a parameter to the system and, on the basis of the resulting successive observations of the mean-square error, finally finds the value of the parameter giving the minimum error. The device is also able to follow the change of the system conditions to keep the system optimum. The optimum value of damping ratio of a second-order system with LF Gaussian noise input is also discussed on the basis of the data obtained with the device.**

Introduction

SELF-optimization, in connection with adaptive control, is currently one of the most interesting and important problems in the field of control technique.

This paper describes an experimental device which self-optimizes a control system by means of a logic circuit. It was especially designed to minimize the square error of a control system, by self-adjusting its parameter, to improve its dynamic performances. From this point of view, the purpose of the device is the same as those of the self-adjusting system described by Anderson, *et al.*,[1] and the self-optimizing servo circuits by Nightingale.[2] The idea of using a logic circuit to optimize the system is similar to that realized in OPCON which was developed by the engineers at Westinghouse[3] and which has been applied to actual industrial chemical processes.

The investigation described in this paper was originally intended to obtain a device that most efficiently searches the optimum condition and maintains it, not by approaching the optimum by slowly increasing or decreasing the value of a parameter, but by finding the upper and the lower limits between which the optimum value is held, and then reducing the range between the limits as fast as possible. For the case of one variable, the sequential minimax search for a minimum or a maximum of a unimodal function has already been discovered by Kiefer.[4] The present paper is also concerned with only one variable for self-adjustment, but the way of search embodied in the device is different from Kiefer's. Therefore, it may not be the best one, but it is still believed to be efficient and has a special feature in that the logical operations decide the sign and size of the move as a trial, and the electric circuits giving the trial value are simplified. This way of search may be extended to the case of two or more variables, which is important particularly for practical applications, but at the present stage of investigation its consideration is left to the future.

In an attempt to obtain an optimum dynamic performance of a control system, the mean-square error is periodically sampled and is put into the device as a sequence of observations which will be denoted as y_i. Immediately after each observation, the device decides the sign and size of the next move as a trial to make the next observation nearer the optimum, according to a rule or a strategy, by taking account of the previous trials and observations, and gives a value x_{i+1} of the parameter to the system. The next observation y_{i+1} is then waited for. In these circumstances, each value of y_i is considered to be a result of the most recent trial made by the device by giving a value x_i of the parameter on the basis of the previous trials and observations. Thus, the device makes successive trials, observes the successive results, and finally finds the optimum value of the parameter. It is also able to follow the change of the system conditions, if any, and to keep the system optimum by continued trials to change the parameter all the time.

An experiment to minimize the mean-square error of a second-order system with a stationary random input was made with a device constructed in this principle. If the error of a control system is considered to be of a stationary random type, as is so in some process control systems, the device could be applied to the system. Also, the device could be applied to a process whose output is desired to be maximized by adjusting or controlling an input variable. No consideration has, however, been made as to the capability of self-optimization of a control system in which no consistent error response to a parameter change is obtained because of an irregular change of the normal operating signals.

* Received by the PGAC, June 24, 1960; revised manuscript received, February 6, 1961.

† University of Tokyo, Tokyo, Japan. Formerly at School of Elec. Engrg., Cornell University, Ithaca, N. Y.

[1] G. W. Anderson, J. A. Aseltine, A. R. Mancini, and C. W. Sarture, "A self-adjusting system for optimum dynamic performance," 1958 IRE National Convention Record, pt. 4, pp. 182–190.

[2] J. M. Nightingale, "Self-optimizing servo circuits," *Machine Design*, vol. 32, pp. 139–143; January 7, 1960.

[3] D. A. Burt, "An optimizing control for the process industries," ASME Rept. No. 59-AUT-4; May, 1959.
J. W. Bernard and F. J. Soderquist, "Dow evaluates optimizing control," *Control Engrg.*, vol. 6, pp. 124–128; November, 1959.
R. Hooke, "Control by automatic experimentation," *Chem. Engrg.*, vol. 64, pp. 284–286; June, 1957.

[4] J. Kiefer, "Sequential minimax search for a maximum," *Proc. Am. Math. Soc.*, vol. 4, pp. 502–506, June, 1953; "Optimum sequential search and approximation methods under minimum regularity assumptions," *J. Soc. Indust. and Appl. Math.*, vol. 5, pp. 105–136, September, 1957.

Strategy

The term "strategy" will be used for the determination of what change is to be given to the value of x to minimize the error of the system; the word "move" will be used for the change itself of the value x at each trial.

In order to minimize the observation by successive trials, two rules were taken: one is about the sign, the other about the size of the move.

A. Rule of the Sign of Move

If the observation y_i is decreased in comparison with the previous one y_{i-1} ($\Delta y = y_i - y_{i-1} < 0$) as a result of either the positive or negative sign of move ($\Delta x = x_i - x_{i-1} \lesseqgtr 0$), this trial is considered successful and the same sign of the move as this trial is taken as the next trial. On the other hand, if the observation y_i is increased ($\Delta y > 0$), this trial is considered a failure and the opposite sign of the move is taken as the next trial.

This rule can be expressed in a simpler form, if logical variables X_i and Y_i are used to represent the signs of change of x and y. It must be emphasized, however, that the use of logical variables in the following is only a means of simply and clearly expressing the rules, or for convenience of logical design of the circuits. The Boolean algebra approach of the optimum search problem is not intended. Now

$$X_i = \begin{cases} 1 & \text{when } \Delta x_i (= x_i - x_{i-1}) > 0 \\ 0 & \text{when } \Delta x_i \leq 0 \end{cases}$$

$$Y_i = \begin{cases} 1 & \text{when } \Delta y_i (= y_i - y_{i-1}) > 0 \\ 0 & \text{when } \Delta y_i \leq 0. \end{cases}$$

The variable X_{i+1}, which represents the sign of the next move, should then be shown as in Table I and may be written in the following algebraic form:

$$X_{i+1} = X_i' Y_i + X_i Y_i' \qquad (1)$$

TABLE I

X_i \ Y_i	0	1
0	0	1
1	1	0

B. Rule of the Size of Move

Eight sizes 1, 2, 4, 8, 16, 32, 64, and 128 are prepared for use, and each of these can be specified by an integer n so that 2^n will be the size. Three successive variables Y_{i-2}, Y_{i-1}, and Y_i, concerning the present and the previous changes of observation, determine the next size of move n_{i+1}. Among eight possible sequences made by these three variables, only at the sequences of 010 and 101 is the next size made one half of the previous one. This is shown in Table II. On the other hand, those of 000 and 111 double it. The other four sequences 001, 011, 100, and 110 cause the size to remain the same. These changes of size should not be made until at least three moves in the same size are taken after one change has occurred.

TABLE II

Y_{i-2}	Y_{i-1}	Y_i	n_{i+1}	R	L
0	0	0	n_i+1	1	1
0	0	1	n_i	0	0
0	1	0	n_i-1	1	0
0	1	1	n_i	0	0
1	0	0	n_i	0	0
1	0	1	n_i-1	1	0
1	1	0	n_i	0	0
1	1	1	n_i+1	1	1

A short explanation of this rule follows. At the obtained sequences 010 and 101, the device has already found the minimum existing between two certain limits because the observed succession is down, up, down or up, down, up, around the minimum. The move of the next trial then should be in a smaller size for the more precise determination of the minimum. The sequence 01 already appears to give sufficient information of the existence of a minimum between two certain limits. It was found though, that to change the size after one more zero, in other words, after the sequence 010 has been observed, is a better strategy if one takes into account the error due to observation and fluctuation of the system response itself, in order to make sure of the existence of the minimum.

At the sequence 000, the successive trials have been successful, but the minimum between two limits has not yet been found. A larger size of move, then, is needed as a trial in order to get the other side of the minimum quickly. This circumstance occurs when the minimum position is changing and the device has to follow up the minimum. The sequence 111 also occurs in a similar circumstance.

Whether or not the size is to be changed, and whether or not the size should be doubled or be halved, can be represented by two logical variables R and L as shown in Table II. They can be written in algebraic expressions as

$$R = Y_i Y_{i-2} + Y_i' Y_{i-2}' \qquad (2)$$

$$L = Y_i Y_{i-1} Y_{i-2} + Y_i' Y_{i-1}' Y_{i-2}'. \qquad (3)$$

To mechanize these logics, the actual device was provided with an eight-digit binary reversible counter functioning as an output register to which one digit was added or subtracted according to the variable X_{i+1} at the specified place representing n_i. The place was made to shift to the right by one when $R=1$ and $L=0$, resulting in n_i-1, and to the left by one when $R=1$ and $L=1$, resulting in n_i+1. This easy mechanizability is the special feature of the strategy and actually made the logic circuit simple. The principle may be applied to other devices such as digital controllers,

Experimental Arrangements

The general arrangement of the self-optimizing device made is shown in the block diagram of Fig. 1. Each block will be explained.

A. A-D Converters (Fig. 2)

A binary counter consisting of eight flip-flops counts clock pulses supplied by a multivibrator (170 cps). The plate of one of the tubes of each flip-flop is connected to a cathode follower whose output is clipped by diode circuits to the two exact values of zero and +35 volts, which correspond to the digits 0 and 1 of the flip-flop. Thus, obtained outputs of the eight flip-flops are applied to a set of resistors (0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, and 6.4 MΩ). The total current through the resistors is proportional to the content of the counter, and thus generated stepwise; increasing current or voltage is compared with the negative voltage to be measured by an amplifier whose output stops the clock pulses passing through a gate to the counter, when the generated voltage just exceeds the voltage to be measured. This simple but relatively low-speed converter was enough for the purpose.

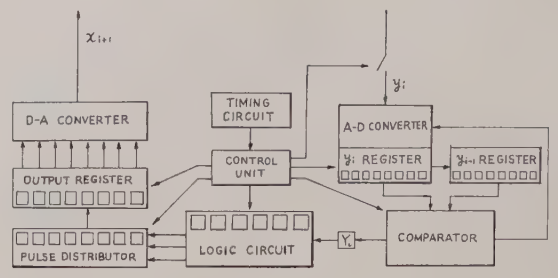


Fig. 1—Block diagram of the self-optimizing device.

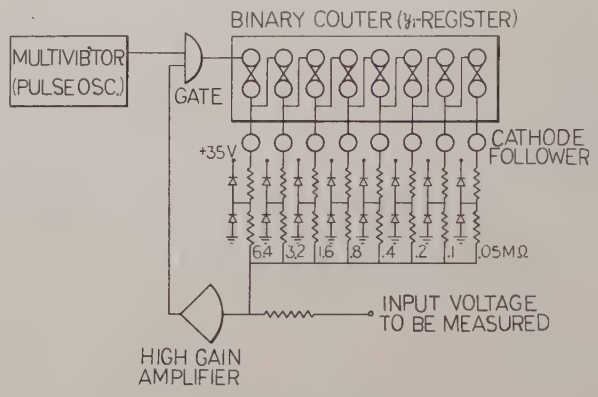


Fig. 2—A-D converter.

B. Comparator of Two Successive Measurements

The binary counter of the A-D converter also functions as a register of the present measurement y_i. Another binary counter also consisting of eight flip-flops was provided as a register of the previous measurement y_{i-1}. These two registers are shown in Fig. 1 as y_i register and y_{i-1} register. A comparator consisting of cascaded gates is provided which gives an output when the contents of the two registers just coincide with each other. If the comparator gives an output during a measurement, the present measurement y_i is larger than the previous one, y_{i-1}. Therefore, this output can be used to actuate a sensitive relay which represents the variable Y_i equal to 1. If it does not, the relay does not operate and the variable Y_i is equal to zero. After the measurement is finished, the content of the first register is transferred to the second register. For this operation, the comparator is again used in such a way that clock pulses allowed to flow into the second register are stopped from passing through a gate by a signal given by the comparator when the contents of the two registers just coincide. At the beginning of the next cycle, the content of the first register is reset, and the new measurement now begins.

C. Logic Circuit, Pulse Distributer and Output Register

The circuits previously mentioned are all elctronic, but the circuits which will be described are made up of groups of relays. The operation of a polar relay representing the variable Y_i is transferred to another relay, designated also as Y_i, in the logic circuit. This one bit of information concerning the increase or the decrease of the system error as a result of trial will be stored until two other successive observations have been made. The variable X_i representing the sign of the previous move is also stored in the circuit. Therefore, four variables in all, Y_{i-2}, Y_{i-1}, Y_i, and X_i, are available to decide the next move. The logic circuit gives a decision following the rule expressed in (1)–(3). It was designed in the usual way with seven relays. The operations of the relays representing the variable X_{i+1}, R and L give instructions to a pulse distributor as to which place on the output register a pulse is to be given. The pulse distributor itself is a reversible ring counter consisting of eight relays. Whether or not the operated state which is stationed on a relay is to remain on the same relay, or move to the next, to the right side, or to the left side, is determined by the variables R and L. Thus, through a determined relay of the pulse distributer, a pulse is sent to the corresponding place on the output register, where an addition or subtraction of one to its content, according to the variable X_{i+1}, takes place. The output register is itself an eight-digit reversible binary counter.

D. Control Unit

The time sequence of operation of all these relays is controlled by a control unit consisting of a set of three main and four auxiliary relays. A start signal causes the released state of the three relays to become unstable. One of them will operate, but this state of combination is still unstable. Therefore, one of the other ones will operate. Thus, the three relays will take every possible eight combination in a definite order. All of them except the final one are unstable, though, and the relays will reach and stop at the final state. During this sequence,

the contacts of the relays offer selectable sets of operating sequences of the other relays and the electronic circuits.

E. D-A Converter

The D-A converter which gives the output value x_{i+1} to the control system consists simply of a set of eight resistors (0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, and 6.4 MΩ) connected in parallel, each of which has a contact of the corresponding relay of the output register in series. If the contact is closed, a current flows through the resistor, but if open, it does not. The summed current represents a value in an analog form corresponding to the digital content of the output resistor. In actual experiments, these resistors in parallel play the part of an input resistor of the analog computer.

F. Timing Circuit

The data from the system are to be periodically sampled, and every operation of the machine is to cycle in a definite manner. A timing circuit is provided for this which gives a pulse at intervals of a definite period. The circuit itself consists of a multivibrator and a four-digit binary counter. With the aid of gate circuits, during the period when all the digits of the counter are one, a positive output voltage is obtained from this circuit which actuates a relay giving the start signal. Actually 45 sec was used as the period of the cycle.

Now the arrangement seems to have a rather large amount of hardware, considering that only one parameter is adjusted for self-optimization. To make the arrangement function properly, however, this amount was demanded. At the start of the experiment, two condensers with an amplifier of an analog computer were used, in place of the A-D converter and the y_i and y_{i+1} electronic registers, to store the present and the previous observations. Also, a high-gain amplifier was used for the comparator, but an unavoidable minute difference between their capacitances and a drift of the comparator gave some trouble. It was found that the measurements of a performance parameter had to be made with an identical measuring unit to obtain a consistent series of observations, and that the data had to be digital in order to be stored and compared without ambiguousness.

Experiment to Minimize the Mean-Square Error

In order to verify the validity of the method, an elementary problem of minimization was taken which requires the minimizing of the error of a second-order system with LF Gaussian noise input by adjusting the damping ratio ζ. The system was simulated on an analog computer.

The equation can be written in the form

$$\frac{d^2z}{dt^2} + 2\zeta\omega_n \frac{dz}{dt} + \omega_n{}^2 z = \omega_n{}^2 v(t), \qquad (4)$$

where $v(t)$ is the input, $z(t)$ is the output, and ω_n the natural angular frequency. This experiment is worthwhile, though elementary, because the dynamic behavior of many complicated control systems can often be approximated by this simple mathematical model. Furthermore, the equation describes the dynamic behavior of most of the indicating and recording instruments such as a mirror galvanometer or a pen-writing oscillograph recorder. There has been a lot of discussion about the optimum value of ζ. Critical damping, in other words, $\zeta = 1$, is sometimes recommended, and at other times the value of $0.707 = 1/\sqrt{2}$ is said to be an optimum. These discussions are, in many cases, based on the indicial or the frequency response to a step or a sinusoidal input.

In the present experiment, a LF Gaussian noise, which was generated by a generator making use of HF noise of a thyratron discharge current and passed through an appropriate low-pass filter, was used as the input to the system. The power spectrum was then given, but the waveform itself was very irregular. In order to discuss the dynamic behavior of an instrument, this type of signal may be more suitable because every instrument has to indicate faithfully or has to record nonpredictable changes of the signal. An active filter whose transfer function is given by $K/(1+Ts)$ was used as a noise filter. Therefore, the power density spectrum of the input to the system was of the type $1/(1+T^2\omega^2)$. The optimum value of ζ is to be 0.5, as was the result when the mean-square error was evaluated (see Appendix).

Fig. 3 shows the block diagram of the experimental arrangement to determine the optimum value of ζ giving a minimum mean-square error on an analog computer by using a diode function generator as the squaring circuit and a dead-time circuit to give a delay to the input. Before the full automatic operation was made with the self-optimizing device, an experiment was performed to determine which value of ζ would give the minimum by changing it manually. Fig. 4 shows one of the results. The agreement with the evaluated value $\frac{1}{2}$ is fairly good. On this determination, three observations made under the same condition were averaged. Actually, the time average of the square error was fluctuating. Theoretically, it would not vary but it did because the time needed to obtain an average in practice is not infinite. The fluctuation of samples of the time average has already been discussed by many authorities for the last thirty years, in connection with the Brownian movement which gives the natural limit of measurement. According to this theory,[5] the fluctuation of the finite-time average is inversely proportional to the time of the average. Therefore, the longer the time, the smaller the fluctuation. In the present device, 45 sec was used for the time, while the time constant of the noise filter was $T = 0.04$ sec, and the natural frequency of the system was 80 cps.

[5] F. Zernike, "Die Brownsche grenze Für Beobachtungsreihen," *Z. Phys.*, vol. 79, pp. 516–528; December, 1932.

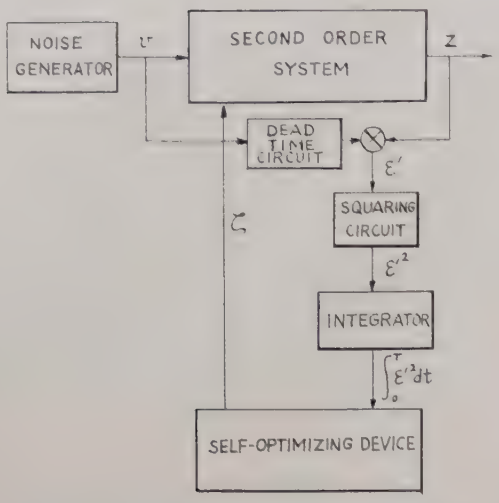


Fig. 3—Experimental arrangement consisting of an analog computer, noise generator, and the self-optimizing device.

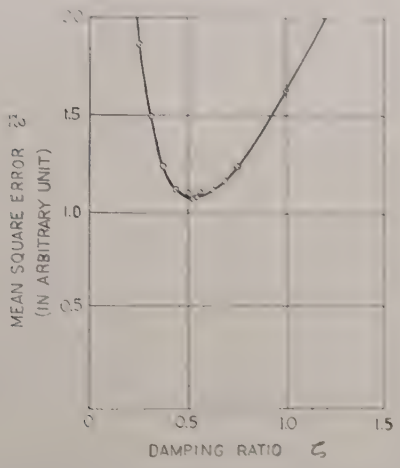


Fig. 4—Experimental determination of the relationship between the mean-square error and the damping ratio.

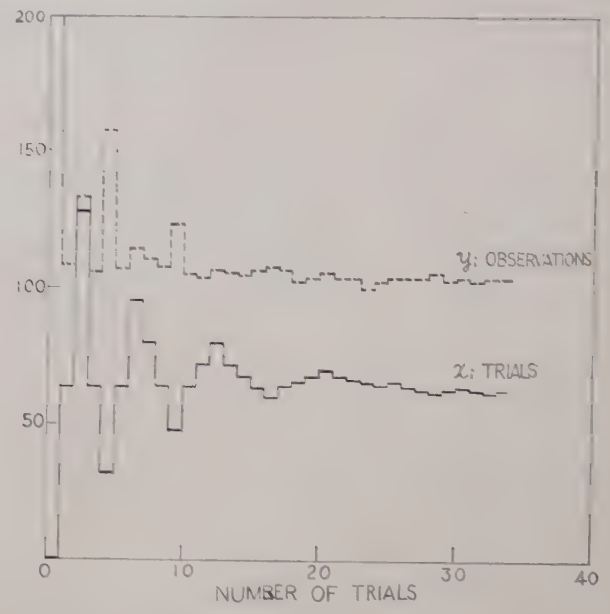


Fig. 5—Automatic search of the minimum error.

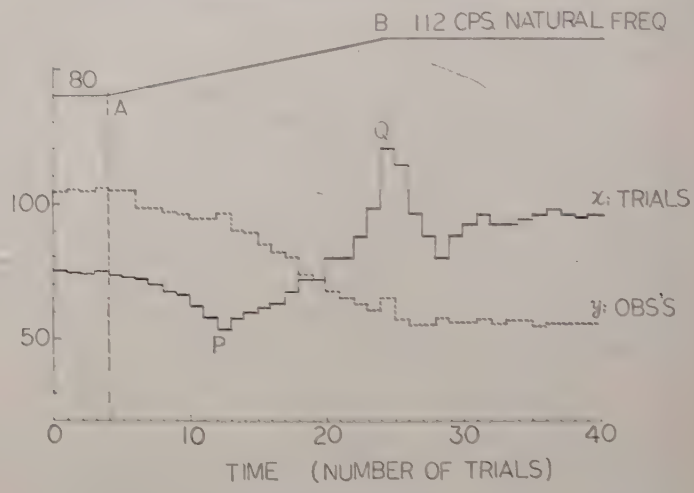


Fig. 6—Follow-up of the changing condition of the system.

The device was instructed to find the minimum value and also to continue to maintain the optimum condition. It was not difficult to set the initial conditions on the machine. After the switch-in, all of the operations were performed automatically.

Some of the results obtained are shown in Figs. 5 and 6. Fig. 5 shows how the device came to find the minimum value by making trials. The x_i's correspond to the ζ values taken as trials, and the y_i's correspond to the averages of the mean-square error observed. In Fig. 6, an example is given of the device following up the changing condition of the system which was caused by the change of the natural frequency of the system from point A to point B. The corresponding trials made by the device and the successive observed values of the mean-square error are also shown in Fig. 6. It is to be noted that the final optimum value of ζ corresponding to the higher natural frequency is larger than the value corresponding to the lower frequency. Therefore, it appears that the optimum value of ζ should increase all the time with the increase of the natural frequency. Nevertheless, the device was actually trying to make it smaller until reaching the point P. This is due to the fact that the observations were decreased even by those trials. At point P, however, the device became aware of the proper step and began to try to make it larger. Thus, it finally came to find the new stable optimum value of ζ.

Conclusion

By using a logic circuit, A-D and D-A converters, and sequential switching circuits, a self-optimizing device which searches a minimum point and follows up a changing condition was obtained, even though it was in an experimental stage. The circuits may be designed in all electronic or transistorized form on the same principle. It is to be noticed here, however, that the speed of minimizing depends mainly upon the response of the

system to a change of the adjustable parameter or upon the smoothing time needed to measure some performance parameters such as the mean-square error of a system with a random input. Actually, in the experiment, the working time of the device was only $\frac{1}{16}$ of the 45 sec needed to obtain the mean-square error. The logic or the strategy for the trials to optimize still remains a problem to be investigated. One set of rules, though, which are easy to mechanize, has been proposed and this same principle may be applied to other devices having similar purposes.

Appendix

The Optimum Damping Ratio of a Second-Order System With a Random Noise Input

The power density spectrum of the output of a linear filter to the input whose power density spectrum is $G_i(\omega)$ is given by

$$G_o(\omega) = G_i(\omega) \left| Y(j\omega) \right|^2, \tag{5}$$

where $Y(j\omega)$ is the frequency response function of the filter; the power density spectrum of the error, then, is

$$G_\epsilon(\omega) = G_i(\omega) \left| Y(j\omega) - 1 \right|^2 \tag{6}$$

and the mean-square error is

$$\overline{\epsilon^2} = \int_0^\infty G_i(\omega) \left| Y(j\omega) - 1 \right|^2 d\omega. \tag{7}$$

Suppose now that the performance of a recording instrument is going to be discussed. The first requirement to be met is that it reproduces the input waveform as faithfully as possible. The short time lag needed to reproduce the input waveform is not important. So instead of using the error just considered, the output signal is compared with a delayed input, and thus the difference between the output signal and a delayed input is considered to be the error, namely,

$$\epsilon'(t) = z(t) - v(t - \tau). \tag{8}$$

Then, the value of the error will be much smaller. In order to evaluate the mean square of this error in the frequency domain, the formula

$$\overline{\epsilon'^2} = \int_0^\infty G_i(\omega) \left| Y(j\omega) - e^{-j\tau\omega} \right|^2 d\omega \tag{9}$$

can be used, where $e^{-j\tau\omega}$ represents the frequence response function of the dead time τ. The damping ratio which minimizes this expression may give, for instance, the optimum condition for the recording instrument to record a waveform whose power density spectrum is $G_i(\omega)$. Two cases of the spectrum are considered in the following:

Case 1):

$$G_i(\omega) = \begin{cases} 1 & \omega \leqq \alpha\omega_n \\ 0 & \omega > \alpha\omega_n \end{cases} \tag{10}$$

This is the case where the input power spectrum is perfectly uniform up to ω_0, $\alpha = \omega_0/\omega_n$, and is sharply cut off right there. This type of filter for the noise is not physically realizable, but is a simple case that is theoretically feasible. Substituting the frequency response function of the second-order system,

$$Y(j\omega) = \frac{1}{-u^2 + j2\zeta u + 1},$$

where $u = \omega/\omega_n$, and also substituting the time lag expressed in λ, a fraction of the natural period T_n, into (9) gives

$$\overline{\epsilon'^2} = \frac{1}{\alpha} \int_0^\alpha \cdot \frac{(-u^2 + 1 - \cos 2\pi\lambda u)^2 + (2\zeta u - \sin 2\pi\lambda u)^2}{(-u^2 + 1)^2 + 4\zeta^2 u^2} du. \tag{11}$$

This integral can be evaluated in terms of a power series of α, when the frequency range of the input signal is confined to a very low limit in comparison with the natural frequency of the system, as

$$\overline{\epsilon'^2} = P\alpha^2 + Q\alpha^4 + R\alpha^6 + \cdots. \tag{12}$$

The coefficient of α^2 is

$$P = \tfrac{2}{3}(\zeta - \pi\lambda)^2.$$

$\lambda_m = \zeta/\pi$ is taken as the value of λ, the first term will be zero, and the coefficient of α^4 will be

$$Q = \frac{4}{225}\zeta(2\zeta^2 - 1)^2.$$

This term will also be zero, if the value

$$\zeta = \tfrac{1}{2} = 0.707 \tag{13}$$

is taken, with

$$\lambda_m = 0.225. \tag{14}$$

The coefficient of the third term will then be

$$R = \frac{2}{63}. \tag{15}$$

Consequently, as long as the frequency component contained in the input signal is confined to a low range in comparison with the natural period of the system, 0.707 is the optimum value in the case.

Case 2):

$$G_i(\omega) = \frac{1}{1 + T^2\omega^2} \cdot$$

This is the case where the noise passes through a filter whose transfer function is $1/(1+Ts)$. The mean-square error is

$$\overline{\epsilon'^2} = \int_0^\infty \frac{1}{1 + \mu^2 u^2} \cdot \frac{(-u^2 + 1 - \cos 2\pi\lambda u)^2 + (2\zeta u - \sin 2\pi\lambda u)^2}{(-u^2 + 1)^2 + 4\zeta^2 u^2} du, \quad (16)$$

where

$$\mu = T\omega_n.$$

This integral can be evaluated by applying the method of residue. The result is again expanded into a power series of $1/\mu$, and will then be

$$\overline{\epsilon'^2} = S\frac{1}{\mu^2} + 0\left(\frac{1}{\mu^3}\right) \cdot \quad (17)$$

The coefficient of the first term is

$$S = \frac{\pi(1 - 4\zeta^2)}{4\zeta} + \frac{\pi}{\sqrt{1 - \zeta^2}} e^{-2\pi\lambda\zeta} \cos(2\pi\lambda\sqrt{1 - \zeta^2} + \phi),$$

where

$$\phi = \tan^{-1}\frac{1 - 2\zeta^2}{2\zeta\sqrt{1 - \zeta^2}} \cdot$$

The condition to make this value zero is given by

$$\zeta = \frac{1}{2}, \quad (18)$$

$$2\pi\lambda\sqrt{1 - \zeta^2} + \phi = \frac{\pi}{2} \cdot \quad (19)$$

From the latter equation $\lambda = 0.193$ (by taking the value of $\zeta = \frac{1}{2}$) is obtained. Therefore, in this case, the optimum damping ratio is $\frac{1}{2}$ insofar as the first term is concerned.

If the value of λ is taken to be equal to zero, or if the error is to be considered in the original sense,

$$S = \frac{\pi(1 + 4\zeta^2)}{4\zeta} \quad (20)$$

which will be minimum, also when

$$\zeta = \tfrac{1}{2}. \quad (21)$$

Acknowledgment

The author wishes to thank Prof. W. E. Meserve for providing the opportunity and facilities for conducting this research at the School of Electrical Engineering, Cornell University, Ithaca, N. Y. and also for offering encouragement and many useful suggestions during the work. The author also wishes to thank the other staff members of the School, especially T. V. McCarthy for constructing the noise generator and Mrs. M. W. Miller for typing the manuscript.

Stability Conditions of Pulse-Width-Modulated Systems Through the Second Method of Lyapunov*

T. T. KADOTA† AND H. C. BOURNE, JR.‡, SENIOR MEMBER, IRE

***Summary*—PWM systems contain inherent nonlinearities which arise from their modulation scheme. Thus, for a legitimate study of stability, such systems must be treated as nonlinear sampled-data systems without initially resorting to linear approximations. For a nonlinear system whose dynamic behavior is described by a set of first-order difference equations, one of the theorems in the second method of Lyapunov gives, as a sufficient condition for asymptotic stability in the large, the existence in the whole space of a positive-definite Lyapunov's function V, whose difference ΔV is negative definite. Hence, by choosing a positive-definite quadratic form as V, the sufficient condition is reduced to the negative-definiteness in the whole space of ΔV. Upon this basis, a systematic procedure of obtaining analytically a sufficient condition for asymptotic stability in the large is developed for various types of PWM systems; the condition is stated as the negativeness of all the eigenvalues of three matrices associated with the PWM system.**

* Received by the PGAC, July 22, 1960; revised manuscript received, March 1, 1961.

† Bell Telephone Labs., Inc., Whippany, N. J. Formerly at University of California, Berkeley, Calif.

‡ University of California, Berkeley, Calif.

Introduction

SINCE the beginning of the last decade, PAM systems, commonly referred to as "sampled-data systems," have been the subject of extensive study. Not until recently, however, has a systematic analysis been attempted on PWM systems in which sampled

data are transmitted in the form of the widths rather than the amplitudes of pulses. This fact may be attributed to the mathematical difficulty involved, since PWM systems contain inherent nonlinearities arising from the modulation scheme, while the PAM systems are basically linear. To the authors' knowledge, the first paper on the PWM system was that by Nease[1] who developed an approximate analysis and design procedure based on two types of linearization depending upon the magnitude of the input to the pulse-width modulator. Andeen[2] presented an approximate analysis by replacing the pulse-width modulator with the equivalent pulse-amplitude modulator with restriction on the magnitude of the input and the sampling period. Polak[3] treated the stability problem without resorting to linearization although his scheme, based on determination of limit cycles, is fundamentally limited to the first-order systems. Instead of analyzing a given PWM system, Nelson[4] viewed such a system as the "pulse-width control" of the sampled-data system, and developed an optimization procedure through proper selection of a PWM "control function."

This paper is concerned with the stability of the PWM system, or more precisely, with obtaining sufficient conditions for stability in terms of given parameters of the system. The general configuration of the system to be considered is shown in Fig. 1, where the linear plant is assumed to be time-invariant. Fig. 2 shows various types of PWM: lead, lag, and lag-delay-integrator;[5] the sampling period T is constant and the width of the pulse is proportional, until it reaches T, to the absolute value of 1) the sampled error for the lead- and lag-type, and 2) the integral of the error during the sampling period for the lag-delay-integrator-type.

For a legitimate study of stability, the PWM system must be treated as a nonlinear system without resorting initially to linear approximation, because 1) the type of stability required is that any initial disturbances around a desired steady state, regardless of their magnitudes, must eventually vanish, thus prohibiting a local linearization around the steady state; and 2) conclusions obtained through linear approximation are invalid in general, if not meaningless, for the original

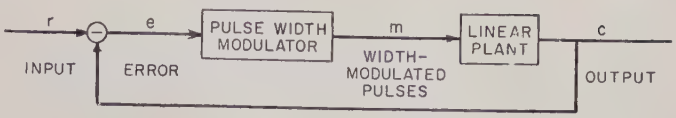


Fig. 1—Schematic diagram of typical PWM systems.

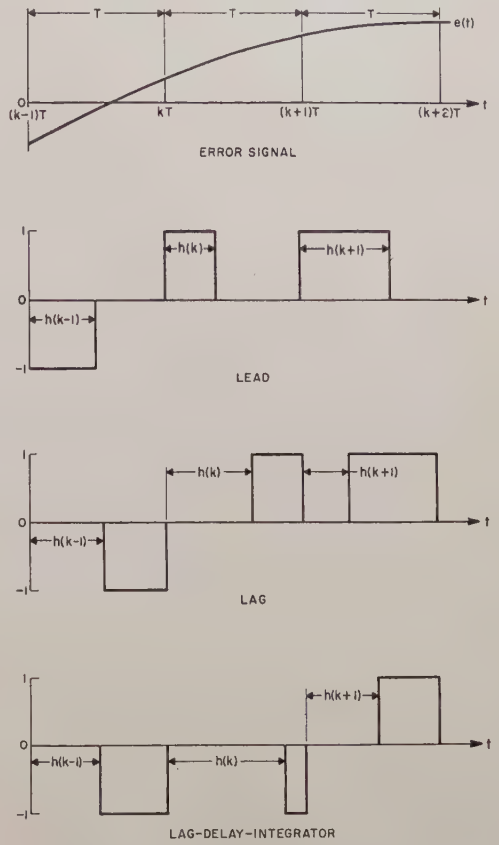


Fig. 2—Various types of PWM.

nonlinear system. Until recently, the two major tools available for solving stability problems were the "describing function" method[7,8] and the phase-plane technique.[8,9] The former is an extension of the frequency response method for linear time-invariant systems to nonlinear systems, and the concept of stability used is essentially that of linear systems, which is, in general, too inadequate a concept to treat nonlinear systems encountered in practice. On the other hand, the phase-plane technique is far more sound in its conceptual structure, and it would be applicable to the PWM system without resorting to linear approximation. However, its use is primarily restricted to second-order systems.

In recent years considerable attention has been paid to a new approach, known as the second method of

[1] R. F. Nease, "Analysis and Design of Nonlinear Sampled-Data Control Systems," Mass. Inst. Tech., Cambridge, Tech. Rept.; June, 1957.

[2] R. E. Andeen, "Analysis of pulse duration sampled-data systems with linear elements," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-5, pp. 306–313; September, 1960.

[3] E. Polak, "Stability Criteria in Pulse Width Modulated Sampled-Data Systems," M.S. thesis, Dept. of Elec. Engrg., University of California, Berkeley; September, 1959.

[4] W. L. Nelson, "Pulse width control of sampled-data systems," *Trans. ASME*, ser. D, vol. 83, pp. 65–76; 1961. ("Optimal control methods for on-off sampling systems," to be published.)

[5] The lead- and the lag-type modulation are the simplest types commonly considered in the literature.[1–4] The lag-delay-integrator modulation appears when a magnetic amplifier is used as the pulse-width modulator.[6]

[6] T. T. Kadota, "Linear Analysis of Control Systems with Magnetic Amplifiers," Electronics Res. Lab., University of California, Berkeley, Tech. Rept., Ser. No. 60, Issue No. 216; March, 1960.

[7] R. J. Kochenburger, "A frequency response method for analyzing and synthesizing contractor servomechanisms," *Trans. AIEE*, vol. 69 (*Commun. and Electronics*), pp. 270–284; 1950.

[8] J. G. Truxal, "Automatic Feedback Control System Synthesis," McGraw-Hill Book Co., Inc., New York, N. Y.; 1955.

[9] A. A. Andronow and C. W. Chaikin, "Theory of Oscillations," Princeton University Press, Princeton, N. J.; 1949.

Lyapunov,[10-12] which is a method of investigating stability of systems of ordinary differential equations without actually solving the equations. The method was originally formulated for systems of differential equations, but has been extended to systems of difference equations. For its rigor in theoretical structure and generality of application, the second method of Lyapunov seems, to the authors' limited knowledge, the most promising approach for the stability study of the PWM system. Consequently, the extension of the method to systems of difference equations, because of the discrete nature of the PWM system, forms the theoretical basis of the material of this paper.

Mathematical Background: Theorem in Second Method of Lyapunov

The assumption of time invariance of the linear plant in Fig. 1 implies that the dynamical behavior of such a PWM system can be described by n first-order autonomous normal difference equations, or by the corresponding vector equation in an n-dimensional Euclidean space, as we shall derive it in the next section. In this space, often called a state space,[12] a solution of the equation is represented by a "state vector," whose components correspond to the output and its $n-1$ derivatives (or limiting values) of the PWM system at sampling instants.[4,12]

It is well known that the question of stability of an arbitrary solution can be reduced, at least formally, to that of the trivial solution of the variational system of difference equations relative to the original solution.[10,13] Geometrically speaking, this trivial solution corresponds to the origin, which is at equilibrium, of the new state space corresponding to the variational system. Notions of stability as well as the second method of Lyapunov have been well established in the literature,[10-13] and we are primarily concerned with asymptotic stability in the large of the trivial solution. In the case of autonomous systems, the definition of asymptotic stability in the large essentially amounts to the following: that any perturbation about the origin, the equilibrium point, regardless of the initial magnitude, is bounded for all time and vanishes eventually.

The mathematical basis for our method of attack is a theorem in the second method of Lyapunov, which gives a sufficient condition for asymptotic stability in the large in terms of definiteness of a "Lyapunov's function" and its difference. Let

$$x(k+1) = f[x(k)] \tag{1}$$

be the vector difference equation in a state space, which describes a given dynamical (discrete in time) system, where k is an integer. We make necessary assumptions on $f(x)$ to assure existence and uniqueness of a solution with an arbitrary initial condition, including the one that $f(0)=0$ which implies existence of the trivial solution $x(k)=0$. A Lyapunov's function $V(x)$ of (1) is a real scalar continuous function of x with $V(0)=0$. The difference (with respect to k) of V is defined as

$$\Delta V = V[f(x)] - V(x), \tag{2}$$

which is the combination of (1) and

$$\Delta V = V[x(k+1)] - V[x(k)].$$

Theorem[14]

If there exists in the whole (state) space a function $V(x)$ which is definite and has the property that $|V(x)| \to \infty$ as $|x| \to \infty$, and if ΔV is also a definite function whose sign is contrary to that of V, then the solution $x(k)=0$ of (1) is asymptotically stable in the large.

Stability Conditions of PWM Systems

On the basis of the preceding theorem, we shall develop the generalized method for obtaining the sufficient conditions for asymptotic stability in the large of the zero steady-state output $c(t)=0$[15] of the two types of PWM systems as shown in Fig. 2. Observe first that the asymptotic stability in the large of $c(kT)=0$ is necessary for that of $c(t)=0$, and it is a trivial matter to establish the latter from the former. Thus, it suffices to consider the asymptotic stability in the large of $c(kT)=0$ only, which leads to examination of the system of difference equations describing $c(kT)$. In anticipation of the use of the preceding theorem, we shall derive a set of the first-order difference equations rather than a single higher-order difference equation. According to the theorem, if we choose a positive-definite quadratic form as a Lyapunov's function V of the system of difference equations in question, then the sufficient condition for asymptotic stability in the large of the trivial solution is the negative-definiteness of ΔV in the whole space. Hence, if we express the sufficient condition for the negative-definiteness of ΔV in the whole space in terms of the given parameters of the PWM system, we shall have obtained the sufficient condition for asymptotic stability in the large of $c(kT)=0$.

In the remainder of this section, we shall present a detailed analytical procedure for obtaining the sufficient condition from a given PWM and a transfer function of the nth-order linear plant.

[10] L. Cesari, "Asymptotic Behaviors and Stability Problems in Ordinary Differential Equations," Springer-Verlag, Berlin, Germany; 1959.

[11] I. G. Malkin, "Theory of Stability of Motion," U. S. Dept. of Commerce, AEC Translation 3352; 1958.

[12] R. E. Kalman and J. E. Bertram, "Control system analysis and design via the 'second method' of Lyapunov: I. Continuous-time system, and II. Discrete-time systems," *Trans. ASME, J. Basic Engrg.*, ser. D, vol. 82, pp. 371–400; June, 1960.

[13] T. T. Kadota, "Asymptotic Stability of Some Nonlinear Feedback Systems," Electronics Res. Lab., University of California, Berkeley, Tech. Rept., Ser. No. 60, Issue No. 264; January, 1960.

[14] A proof of the theorem is given in Appendix I.

[15] In the case of a nonzero steady-state output, the variational system must be derived before application of the method presented here. However, for an ordinary nonzero steady-state output, such as a constant output in a regulating system, derivation of the variational system is a trivial matter.

A. With Lead-Type PWM

Suppose the input to the linear plant is described by

$$m(t)=\{H(t-kT)-H[t-kT-h(k)]\}\operatorname{sgn} e(kT),\ kT<t<(k+1)T, \tag{3}$$

where H and sgn are a Heaviside unit function and a signum function, respectively, which are defined as

$$H(t)=\begin{cases}1 \text{ for } t>0,\\ 0 \text{ for } t<0;\end{cases}\qquad \operatorname{sgn} x=\begin{cases}1 \text{ for } x>0,\\ -1 \text{ for } x<0;\end{cases}$$

and $h(k)$ is the width of the kth pulse which is expressed as

$$h(k)=T \operatorname{sat}\frac{d}{T}\,|\,e(kT)\,|\,,$$

where d is a positive constant and the saturation function is defined as

$$\operatorname{sat} x=\begin{cases}1 \text{ for } & x & >1,\\ x \text{ for } & |\,x\,| & \leq 1,\\ -1 \text{ for } & x & <-1.\end{cases}$$

The transfer function of the linear plant is given by

$$G(s)=b_0\frac{(s-\beta_1)(s-\beta_2)\cdots(s-\beta_j)}{(s-\alpha_1)(s-\alpha_2)\cdots(s-\alpha_n)},\qquad n>j, \tag{4}$$

where b_0 is a positive constant, and α's and β's are real and distinct.[16] Then, the differential equation describing this PWM system becomes

$$c^{(n)}+a_1c^{(n-1)}+\cdots+a_nc=b_0m^{(j)}+b_1m^{(j-1)}+\cdots+b_jm,$$

where $c^{(i)}=d^ic/dt^i$, and a's and b's are defined by

$$(s-\alpha_1)(s-\alpha_2)\cdots(s-\alpha_n)=s^n+a_1s^{n-1}+\cdots+a_n,$$

$$b_0(s-\beta_1)(s-\beta_2)\cdots(s-\beta_j)=b_0s^j+b_1s^{j-1}+\cdots+b_j,$$

and $c^{(n)}$, $c^{(n-1)}$, $\cdots\ddot{c}$; $m^{(j)}$, $m^{(j-1)}$, $\cdots$, $\dot{m}$ are symbolic functions.[17] Through the following standard transformation:

$$c=x_1,\ \dot{c}=x_2,\ \cdots,\ c^{(n-1)}=x_n,\ \text{and}\ \sum_{r=0}^{j}b_rm^{(j-r)}=u_n,$$

the above nth order equation may be transformed into a set of n first-order normal differential equations, which can be written in the matrix form as follows:

$$\dot{x}=Ax+u, \tag{5}$$

where

$$A=\begin{bmatrix}0 & 1 & & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \cdot & \cdot & \cdot & \cdots & \cdot\\ \cdot & \cdot & \cdot & \cdots & \cdot\\ 0 & 0 & & \cdots & 1\\ -a_n & -a_{n-1} & & \cdots & -a_1\end{bmatrix}$$

$$x=\begin{bmatrix}x_1\\ x_2\\ \cdot\\ \cdot\\ x_{n-1}\\ x_n\end{bmatrix},\qquad u=\begin{bmatrix}0\\ 0\\ \cdot\\ \cdot\\ 0\\ u_n\end{bmatrix}.$$

According to the well known theorem on similarity transformation,[17] there exists a real nonsingular matrix P such that $P^{-1}AP=J$ is a diagonal matrix. Suppose we choose for P a matrix whose elements in the first row are all unity, and introduce a change of variables defined by

$$x=Py \tag{6}$$

Then, substitution of (6) into (5) results in the following:

$$\dot{y}=Jy+v, \tag{7}$$

where

$$J=\begin{bmatrix}\alpha_1 & 0 & \cdots & 0\\ 0 & \alpha_2 & \cdots & 0\\ \cdot & & \cdots & \\ \cdot & & \cdots & \\ 0 & 0 & \cdots & \alpha_n\end{bmatrix}\qquad y=\begin{bmatrix}y_1\\ y_2\\ \cdot\\ \cdot\\ y_n\end{bmatrix}$$

$$v=P^{-1}u=\begin{bmatrix}v_1\\ v_2\\ \cdot\\ \cdot\\ v_n\end{bmatrix}.$$

Note that, from (6)

$$x_1=\sum_{s=1}^{n}y_s, \tag{8}$$

and, from (7)

$$v_i(t)=-(P^{-1})_{in}\Bigg[\sum_{r=0}^{j-1}b_r\{\delta^{(j-r-1)}(t-kT)-\delta^{(j-r-1)}[t-kT-h(k)]\}+b_j\{H(t-kT)-H[t-kT-h(k)]\}\Bigg]\cdot\operatorname{sgn}\sum_{s=1}^{n}y_s(kT). \tag{9}$$

[16] The cases where some of the poles are complex conjugate or multiple are treated in Appendix II.

[17] B. Friedman, "Principles and Techniques of Applied Mathematics," John Wiley and Sons, Inc., New York, N. Y.; 1956.

Note that $e = -c = -x_1$.

From the well-known theorem on a linear nonhomogeneous system of differential equations with constant coefficients,[18] the general solution of (7) is

$$y = e^{(t-t_0)J}y(t_0) + \int_{t_0}^{t} e^{(t-\tau)J}v(\tau)d\tau. \tag{10}$$

By substituting into (10) the following two sets of values of t_0 and t:

$$t_0 = kT - 0, \qquad t = kT + h(k) + 0;$$

$$t_0 = kT + h(k) + 0, \qquad t = (k+1)T - 0,$$

we obtain

$$y[kT + h(k)] = e^{h(k)J}y(kT) + \int_{kT-0}^{kT+h(k)+0} e^{[kT+h(k)+0-\tau]J}v(\tau)d\tau,$$

$$y[(k+1)T] = e^{[T-h(k)]J}y[kT + h(k)],$$

where $y(kT)$ and $y[kT+h(k)]$ are used for $y(kT-0)$ and $y[kT+h(k)+0]$, respectively, for notational simplicity. By combining the two equations above,

$$y[(k+1)T] = e^{TJ}y(kT) + e^{[T-h(k)]J}\int_{kT-0}^{kT+h(k)+0} e^{[kT+h(k)+0-\tau]J}v(\tau)d\tau.$$

Upon completion of the integration in the above, with the substitution of (9), we obtain

$$y(k+1) = e^{TJ}[y(k) + w(k)],$$

or equivalently

$$y_i(k+1) = e^{\alpha_i T}[y_i(k) + w_i(k)], \tag{11}$$

where

$$w_i(k) = \Gamma_i \frac{\exp\left[-\alpha_i T \operatorname{sat}\frac{d}{T}\left|\sum_{s=1}^{n} y_s(k)\right|\right] - 1}{\alpha_i} \cdot \operatorname{sgn}\sum_{s=1}^{n} y_s(k),$$

$$\Gamma_i = (P^{-1})_{in}\sum_{r=0}^{j} b_r \alpha_i{}^{j-r},$$

and $i = 1, 2, \cdots, n$, and the sampling period T has been omitted for notational simplicity. Note that the substitution for $h(k)$ has been made through (3).

Suppose we choose as a Lyapunov's function of system (11)

$$V = \sum_{i=1}^{n} \gamma_i y_i{}^2(k), \tag{12}$$

where γ's are arbitrary positive constants. V is obviously positive-definite. Then,

$$\Delta V = \sum_{i=1}^{n} \gamma_i[e^{2\alpha_i T}(y_i + w_i)^2 - y_i{}^2], \tag{13}$$

where $y_i(k)$ and $w_i(k)$ have been abbreviated by y_i and w_i for notational simplicity. Because of the form of V, it is convenient at this stage to introduce the polar coordinates which are defined as follows:

$$\left.\begin{aligned} y_1 &= r\cos\theta_1 &&= r\phi_1,\\ y_2 &= r\sin\theta_1\cos\theta_2 &&= r\phi_2,\\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ y_{n-1} &= r\sin\theta_1\cdots\sin\theta_{n-2}\cos\theta_{n-1} &&= r\phi_{n-1},\\ y_n &= r\sin\theta_1\cdots\sin\theta_{n-2}\sin\theta_{n-1} &&= r\phi_n. \end{aligned}\right\} \tag{14}$$

In addition, we introduce the following notation for later convenience:

$$\sum_{i=1}^{n} \phi_i = \phi.$$

Substitution of (14) into (13) results in

$$\Delta V = r^2 \sum_{i=1}^{n} \gamma_i\left[e^{2\alpha_i T}\left(\phi_i + \frac{w_i}{r}\right)^2 - \phi_i{}^2\right].$$

From the theorem of the preceding section, the sufficient condition for asymptotic stability in the large of $y=0$ of system (11) becomes

$$\sum_{i=1}^{n} \gamma_i\left[e^{2\alpha_i T}\left(\phi_i + \Gamma_i \frac{\exp\left[-\alpha_i T \operatorname{sat}\frac{d}{T} r|\phi|\right] - 1}{\alpha_i r|\phi|}\phi\right)^2 - \phi_i{}^2\right] < 0$$

for all $r \neq 0$ and ϕ_i. (15)

Examination of (15) shows that the function inside the small bracket is a piecewise monotonic function of r. Namely, it is monotonic with respect to r within the following two intervals:

$$0 \leq r \leq \frac{T}{d|\phi|}, \qquad \frac{T}{d|\phi|} \leq r < \infty. \tag{16}$$

Hence, the maximum of the square of the small bracket, and consequently, the maximum of the left-hand side of (15) occurs at either one of the three boundaries, namely, $r=0$, $T/d|\phi|$ and ∞. Therefore, the condition

[18] E. A. Coddington and N. Levinson, "Theory of Ordinary Differential Equations," McGraw-Hill Book Co., Inc., New York, N. Y.; 1955.

(15) is reduced to the following:

$$\left.\begin{aligned}&\sum_{i=1}^{n}\gamma_i[e^{2\alpha_i T}(\phi_i+p_i\phi)^2-\phi_i{}^2]<0,\\&\sum_{i=1}^{n}\gamma_i[e^{2\alpha_i T}(\phi_i+q_i\phi)^2-\phi_i{}^2]<0,\\&\sum_{i=1}^{n}\gamma_i(e^{2\alpha_i T}-1)\phi_i{}^2<0,\end{aligned}\right\}\quad\text{for all }\phi_i,\quad(17)$$

where

$$p_i=-\Gamma_i d,\qquad q_i=\Gamma_i d\frac{e^{-\alpha_i T}-1}{\alpha_i T},\qquad i=1,2,\cdots,n.$$

Since the left-hand sides of the first two inequalities of (17) are in real quadratic forms, they can be expressed as the following scalar products:

$$r^{-2}(y,By),\qquad r^{-2}(y,Cy),$$

where B and C are symmetric matrices. Then, from the property of real quadratic forms,[17] the scalar products (y, By) and (y, Cy) are negative-definite if, and only if, all the eigenvalues of B and C are negative. Thus, the first two conditions of (17) are reduced to all the eigenvalues of the symmetric matrices, B and C, being negative. It is self-evident that the third condition of (17) is reduced to $\alpha_i<0$, $i=1, 2, \cdots, n$; namely, all the eigenvalues of A, the matrix characterizing the linear plant, are negative.[19] Hence, we conclude that $y=0$ of system (11) is asymptotically stable in the large if all the eigenvalues of the matrices A, B and C are negative.

If we recall the earlier substitutions,

$$c=x_1\quad\text{and}\quad x_i(kT)=\sum_{i=1}^{n}y_i(kT),$$

and also the fact that $c(t)$ is an ordinary continuous function of t, implying

$$\lim_{\epsilon\to 0}\sum_{i=1}^{n}y_i(kT-\epsilon)=\sum_{i=1}^{n}y_i(kT),$$

the condition that all the eigenvalues of the matrices A, B and C be negative assures the asymptotic stability in the large of $c(kT)=0$. Then the condition that all the eigenvalues of A be negative assures the asymptotic stability in the large of $c(t)=0$. Since all the elements of the matrices A, B and C are exclusively determined by the given parameters of the PWM system except the arbitrary constants γ's, which may appropriately be determined for a specific problem, the condition that all the eigenvalues of A, B and C be negative is the desired sufficient condition for asymptotic stability in the large of $c(t)=0$ of the PWM system.

[19] The case where one of the eigenvalues of A is zero can be treated also with slight modification. Such a treatment is essentially shown in Part B of this section.

B. With Lag-Delay-Integrator-Type PWM

In the case with the lag-delay-integrator-type PWM, the input to the linear plant is described by

$$m(t)=\{H[t-kT-h(k)]-H[t-(k+1)T]\}\cdot\operatorname{sgn}\int_{(k-1)T}^{kT}e(\tau)d\tau,\qquad kT<t<(k+1)T,\quad(18)$$

where

$$h(k)=T-T\operatorname{sat}\frac{d}{T}\left|\int_{(k-DT)}^{kT}e(\tau)d\tau\right|,$$

and we assume (4) to be the transfer function of the plant. Because of the integral of the error signal $e(\tau)$ in the expression for the pulse width in the above, the appropriate change of variables for this case is the following:

$$\int c\,dt=x_1,\qquad c=x_2,\cdots,\qquad c^{(n-1)}=x_{n+1};$$

$$\sum_{r=0}^{j}b_r m^{(j-r)}=u_{n+1}.$$

Through this transformation, we obtain

$$\dot{x}=Ax+u,\qquad(19)$$

where

$$A=\begin{bmatrix}0&1&0&\cdots&0\\0&0&1&\cdots&0\\\cdot&\cdot&\cdot&\cdots&\\\cdot&\cdot&\cdot&\cdots&\\0&0&0&\cdots&1\\0&-a_n&-a_{n-1}&\cdots&-a_1\end{bmatrix},$$

$$x=\begin{bmatrix}x_1\\x_2\\\cdot\\\cdot\\x_n\\x_{n+1}\end{bmatrix},\qquad u=\begin{bmatrix}0\\0\\\cdot\\\cdot\\0\\u_{n+1}\end{bmatrix}.$$

In order to diagonalize the matrix A, we introduce another change of variables defined by

$$x=Py,\qquad(20)$$

where P is a nonsingular real matrix whose elements in the first row are all unity such that

$$P^{-1}AP=\begin{bmatrix}0&0&\cdot&\cdot&\cdot&0\\0&\alpha_1&\cdot&\cdot&\cdot&0\\\cdot&\cdot&\alpha_2&\cdot&\cdot&\cdot\\\cdot&\cdot&\cdot&\cdot&\cdot&\cdot\\0&0&\cdot&\cdot&\cdot&\alpha_n\end{bmatrix}.$$

Substitution of (20) into (19) results in

$$\dot{y} = Jy + v, \tag{21}$$

where

$$J = P^{-1}AP, \qquad v = P^{-1}u.$$

Note that

$$x_1 = \sum_{s=1}^{n+1} y_s,$$

$$\begin{aligned} v_i(t) = &- (P^{-1})_{i,n+1} \Bigg\{ \sum_{r=0}^{j-1} b_r \{ \delta^{(j-r-1)}[t - kT - h(k)] \\ &- \delta^{(j-r-1)}[t - (k+1)T] \} \\ &+ b_j \{ H[t - kT - h(k)] - H[t - (k+1)T] \} \Bigg\} \\ &\cdot \operatorname{sgn} \sum_{s=1}^{n+1} \{ y_s(kT) - y_s[(k-1)T] \}. \end{aligned} \tag{22}$$

Substitution of the following values of t_0 and t:

$$t_0 = kT + 0, \qquad t = kT + h(k) - 0;$$
$$t_0 = kT + h(k) - 0, \qquad t = (k+1)T + 0,$$

into (10) gives the expression for $y[kT+h(k)]$ and $y[(k+1)]$, which we combine to obtain

$$\begin{aligned} y[(k+1)T] = &\, e^{TJ} y(kT) \\ &+ \int_{kT+h(k)-0}^{(k+1)T+0} e^{[(k+1)T+0-\tau]J} v(\tau) d\tau. \end{aligned} \tag{23}$$

Upon substitution of (22) into (23) and completion of the integration, we obtain

$$y(k+1) = e^{TJ} y(k) + w(k, k-1),$$

or equivalently,

$$\begin{aligned} y_1(k+1) &= y_1(k) + w_1(k, k-1), \\ y_i(k+1) &= e^{\alpha_{i-1}T} y_i(k) + w_i(k, k-1), \end{aligned} \tag{24}$$

where

$$w_1(k, k-1) = - \Gamma_1' T \operatorname{sat} \frac{d}{T} \sum_{s=1}^{n+1} [y_s(k) - y_s(k-1)],$$

$$w_i(k, k-1) = \Gamma_i' \frac{1 - \exp \left\{ \alpha_{i-1} T \operatorname{sat} \frac{d}{T} \left| \sum_{s=1}^{n+1} [y_s(k) - y_s(k-1)] \right| \right\}}{\alpha_{i-1}} \cdot \operatorname{sgn} \sum_{s=1}^{n+1} [y_s(k) - y_s(k-1)],$$

$$\Gamma_1' = (P^{-1})_{1,n+1} b_j, \qquad \Gamma_i' = (P^{-1})_{i,n+1} \sum_{r=0}^{j} b_r \alpha_{i-1}^{j-r},$$

and $i = 2, 3, \cdots, n+1$.

In order to transform the set of $n+1$ second-order difference equations (24) into a set of first-order equations so that the theorem in the preceding section is applicable, we introduce the following change of variables:

$$\begin{aligned} y_1(k-1) &= z_1(k), \\ y_2(k-1) &= z_2(k), \cdots, y_{n+1}(k-1) = z_{n+1}(k), \\ y_1(k) &= z_{n+2}(k), \\ y_2(k) &= z_{n+3}(k), \cdots, y_{n+1}(k) = z_{2n+2}(k). \end{aligned}$$

Then, (24) is transformed into the following set of $2n+2$ first-order difference equations:

$$\left.\begin{aligned} z_1(k+1) &= z_{n+2}(k), \\ z_i(k+1) &= z_{n+1+i}(k), \\ z_{n+2}(k+1) &= z_{n+2}(k) + w_{n+2}'(k), \\ z_{n+1+i}(k+1) &= e^{\alpha_{i-1}T} z_{n+1+i}(k) + w_{n+1+i}'(k), \end{aligned}\right\} \tag{25}$$

where

$$w_{n+2}'(k) = - \Gamma_1' T \operatorname{sat} \frac{d}{T} \sum_{s=1}^{n+1} [z_{n+1+s}(k) - z_s(k)],$$

$$w_{n+1+i}'(k) = \Gamma_i' \frac{1 - \exp \left\{ \alpha_{i-1} T \operatorname{sat} \frac{d}{T} \left| \sum_{s=1}^{n+1} [z_{n+1+s}(k) - z_s(k)] \right| \right\}}{\alpha_{i-1}} \cdot \operatorname{sgn} \sum_{s=1}^{n+1} [z_{n+1+s}(k) - z_s(k)],$$

and $i = 2, 3, \cdots, n+1$.

Suppose we choose as a Lyapunov's function of system (25)

$$V = \gamma_1 z_{n+2}^2(k) + \sum_{i=2}^{n+1} \gamma_i z_{n+1+i}^2(k), \tag{26}$$

where γ_1 and γ_i are arbitrary positive constants and $i = 2, 3, \cdots, n+1$.[20] Then,

$$\begin{aligned} \Delta V = &\, \gamma_1 [(z_{n+2} + w_{n+2}')^2 - z_{n+2}^2] \\ &+ \sum_{i=2}^{n+1} \gamma_i [(e^{\alpha_{i-1}T} z_{n+1+i} + w_{n+1+i}')^2 - z_{n+1+i}^2]. \end{aligned} \tag{27}$$

By introducing the following polar coordinates,

$$\begin{aligned} z_1 &= r \cos \theta_1 &&= r\phi_1, \\ z_2 &= r \sin \theta_1 \cos \theta_2 &&= r\phi_2, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ z_{2n+2} &= r \sin \theta_1 \sin \theta_2 \cdots \sin \theta_{2n+1} &&= r\phi_{2n+2}, \end{aligned}$$

and also the following notations for later convenience,

$$\phi = \sum_{s=1}^{n+1} \phi_s, \qquad \phi' = \sum_{s=1}^{n+1} \phi_{n+1+s},$$

[20] Justification of such a choice of a Lyapunov's function is made in Appendix I.

(27) may be written as

$$\Delta V = r^2\left\{\gamma_1\left[\left(\phi_{n+2} + \frac{w_{n+2}'}{r}\right)^2 - \phi_{n+2}^2\right] + \sum_{i=2}^{n+1}\gamma_i\left[\left(e^{\alpha_{i-1}T}\phi_{n+1+i} + \frac{w_{n+1+i}'}{r}\right)^2 - \phi_{n+1+i}^2\right]\right\}.$$

From the theorem in the preceding section, the sufficient condition for asymptotic stability in the large of $z=0$ of system (25) becomes

$$\gamma_i\left\{\left[\phi_{n+2} - \Gamma_1'\frac{T\,\mathrm{sat}\frac{d}{T}r(\phi'-\phi)}{r}\right]^2 - \phi_{n+2}^2\right\} + \sum_{i=2}^{n+1}\gamma_i\left\{\left[e^{\alpha_{i-1}T}\phi_{n+1+i} + \Gamma_i'\cdot\frac{1-\exp\left[\alpha_{i-1}T\,\mathrm{sat}\frac{d}{T}r|\phi'-\phi|\right]}{\alpha_{i-1}r|\phi'-\phi|}(\phi'-\phi)\right]^2 - \phi_{n+1+i}^2\right\} < 0, \quad \text{for all } r \neq 0, \phi_1', \phi_i, \phi_{n+2}, \phi_{n+1+i}, \tag{28}$$

where $i=2, 3, \cdots, n+1$. By using the argument on a piecewise monotonic function of r, the condition (28) is reduced to the following:

$$\left.\begin{aligned}
&\gamma_1\{[\phi_{n+2} + p_1(\phi'-\phi)]^2 - \phi_{n+2}^2\} \\
&+ \sum_{i=2}^{n+1}\gamma_i\{[e^{\alpha_{i-1}T}\phi_{n+1+i} + p_i(\phi'-\phi)]^2 - \phi_{n+1+i}^2\} < 0, \\
&\gamma_1\{[\phi_{n+2} + q_1(\phi'-\phi)]^2 - \phi_{n+2}^2\} \\
&+ \sum_{i=2}^{n+2}\gamma_i\{[e^{\alpha_{i-1}T}\phi_{n+1+i} + q_i(\phi'-\phi)]^2 - \phi_{n+1+i}^2\} < 0, \\
&\sum_{i=2}^{n+1}\gamma_i(e^{2\alpha_{i-1}T} - 1)\phi_{n+1+i}^2 < 0, \\
&\qquad\qquad \text{for all } \phi_1, \phi_i, \phi_{n+2}, \phi_{n+1+i},
\end{aligned}\right\} \tag{29}$$

where

$$p_1 = -\Gamma_1'd, \qquad p_i = -\Gamma_i'd,$$

$$q_1 = -\Gamma_1'd, \qquad q_i = \Gamma_i'd\frac{1-e^{\alpha_{i-1}T}}{\alpha_{i-1}T},$$

and $i=2, 3, \cdots, n+1$. Note that the left-hand sides of the first two inequalities of (29) are in real quadratic forms, and that the third condition is equivalent to $\alpha_{i-1}<0$.

The remainder of the procedure of obtaining the sufficient condition for asymptotic stability in the large is the same as that in the case with the lead-type PWM.

Conclusions

An analytical procedure of obtaining a sufficient condition for asymptotic stability in the large of the PWM system is developed through the use of one of the theorems of the second method of Lyapunov. In order for the theorem to be applicable, a set of first-order difference equations for the system is derived from a given transfer function of the linear plant and a pulse-width-modulator. By choosing a sum of squares of the coordinates as a positive-definite Lyapunov's function V and then introducing the polar coordinates, the difference of the Lyapunov's function ΔV can be shown to be a piecewise monotonic function of the radius r of the polar coordinates. Hence, the negative-definiteness of ΔV with respect to r is reduced to the negativeness of ΔV at the boundaries of the intervals of r within which ΔV is monotonic. This reduction results, in essence, in changing an exponential form of ΔV into a set of real quadratic forms, which can be expressed as scalar products involving symmetric matrices. Thus, the condition that ΔV be negative-definite in the whole space is finally reduced to the condition that all the eigenvalues of these symmetric matrices be negative, which is the sufficient condition for asymptotic stability in the large of the trivial solution of the set of difference equations for the system. It is a simple matter to deduce from this condition the asymptotic stability in the large of the zero steady-state output of the PWM system.

The condition for asymptotic stability in the large thus obtained is sufficient, but not necessary-and-sufficient. This is an inherent consequence of the use of the theorem, which does not provide a unique way of selecting a Lyapunov's function. Thus, the degree of necessity of the obtained condition depends upon the appropriateness of the choice of a Lyapunov's function for a given system, and the use of a simple Lyapunov's function such as a sum of squares of the coordinates may often give a rather conservative result. One obvious way of improving the necessity of the condition is to choose a more appropriate Lyapunov's function for a given system. However, because of the type of the nonlinearity of the PWM system, the choice of a Lyapunov's function is rather limited if the procedure of obtaining the condition is to be analytically feasible.[21] Another possibility of improvement may be to obtain a condition for instability through the use of the theorems on instability in the second method of Lyapunov, thus setting an upper-bound, so to speak, while the condition for stability sets a lower-bound.

In conclusion, the method presented in this paper provides a systematic procedure for obtaining *analytically* the sufficient condition for asymptotic stability in the large for the general types of PWM systems.[22]

[21] For example, the choice of a sum of squares of the coordinates rather than a more general quadratic form is necessary for the argument on piecewise monotonic functions to be effective.

[22] The method is also applicable to PAM systems with nonlinear gains and some continuous systems with nonlinear gains.[13]

Appendix I

Proof of Theorem and Comment on Choice of Lyapunov's Function

Theorem

If there exists in the whole space a function $V(x)$ which is definite and has the property that $|V(x)| \to \infty$ as $\|x\| \to \infty$, and if ΔV is also a definite function whose sign is contrary to that of V, then the solution $x=0$ of (1) is asymptotically stable in the large.

Proof: Without loss of generality, we may assume that V is positive-definite and ΔV is negative-definite. Then,

$$V(x) > 0 \quad \text{for all } x \neq 0, \quad \text{and} \quad V(0) = 0, \tag{30}$$

$$|V(x)| \to \infty \quad \text{as } \|x\| \to \infty, \tag{31}$$

$$\Delta V(x) < 0 \quad \text{for all } x \neq 0, \quad \text{and} \quad \Delta V(0) = 0. \tag{32}$$

From the uniqueness of a solution of (1) for an arbitrary solution $x(k; k_0, x^0)$ with $x^0 \neq 0$,

$$x(k; k_0, x^0) \neq 0 \qquad \text{for all } k \geq k_0.$$

Hence, from (32),

$$\Delta V[x(k; k_0, x^0)] < 0 \qquad \text{for all } k \geq k_0. \tag{33}$$

Then, from (30) and (33), there exists an $\alpha \geq 0$ such that

$$\lim_{k \to \infty} V[x(k; k_0, x^0)] = \alpha. \tag{34}$$

We shall prove by contradiction that such an α must be zero.

Suppose $\alpha > 0$, then, from (34) and continuity of V in x, there must exist an $a > 0$ such that

$$\|x(k; k_0, x^0)\| > a \qquad \text{for all } k \geq k_0.$$

Then, from (31) and (32), there must exist a $\beta > 0$ such that

$$\Delta V[x(k; k_0, x^0)] \leq -\beta \qquad \text{for all } k \geq k_0.$$

Consequently, we should have the following inequality for all $k \geq k_0$:

$$V[x(k; k_0, x^0)] = V(x^0) + \sum_{k'=k_0}^{k-1} \Delta V[x(k'; k_0, x^0)] \leq V(x^0) - \beta(k - k_0),$$

which is in contradiction with (30) since the right-hand side of the inequality becomes negative for a sufficiently large k. Hence,

$$\lim_{k \to \infty} V[x(k; k_0, x^0)] = 0. \tag{35}$$

Then, from (30),

$$\lim_{k \to \infty} x(k; k_0, x^0) = 0, \tag{36}$$

which proves the theorem.

Comment: In general, a Lyapunov's function V must be a function of all the coordinates of the space. Otherwise, (35) does not necessarily imply (36). Namely, some of the components of $x(k; k_0, x^0)$, of which V is not a function, may not vanish as $k \to \infty$; consequently, the asymptotic stability in the large of $x=0$ cannot be established.

However, as in the case of (25), if

$$\lim_{k \to \infty} x_i(k; k_0, x^0) = 0, \qquad i = 1, 2, \cdots, m; m < n,$$

implies

$$\lim_{k \to \infty} x_i(k; k_0, x^0) = 0, \qquad i = m+1, m+2, \cdots, n,$$

then it is sufficent to choose a V which is a function of x_i, $i=1, 2, \cdots, m$, in order to assure the asymptotic stability in the large of $x=0$.

Appendix II

Cases Where Some of Poles of Linear Plant are Complex or Multiple

A. Complex Poles

With the assumption that the linear plant is a physical system, if some of the poles of the transfer function $G(s)$ of (3) are complex they must appear as pairs of complex conjugate poles. For simplicity, we assume that $G(s)$ has only one pair of complex poles, namely, $\alpha_1 = \alpha_2^* = \xi + j\eta$ were ξ and η are real. In order to obtain a system of real equations corresponding to (11), which is necessary for the introduction of the real polar coordinates (14), the transformation matrix P in (6) is chosen as follows: P is real nonsingular matrix with all the elements in the first row being unity such that $P^{-1}AP = K$ where

$$K = \begin{bmatrix} \xi & \eta & 0 & \cdots & 0 \\ -\eta & \xi & 0 & \cdots & 0 \\ 0 & 0 & \alpha_3 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & \alpha_n \end{bmatrix}.$$

Then, the corresponding equation to (7) becomes

$$\dot{y} = Ky + v, \tag{37}$$

and the relations (8) and (9) remain unchanged. By following the same procedure as in the case of real simple poles, we obtain the following set of n first-order difference equations corresponding to (11):

$$\begin{aligned} y_1(k+1) &= e^{\xi T}[y_1(k)\cos\eta T - y_2(k)\sin\eta T + w_1(k)], \\ y_2(k+1) &= e^{\xi T}[y_1(k)\sin\eta T + y_2(k)\cos\eta T + w_2(k)], \\ y_i(k+1) &= e^{\alpha_i T}[y_i(k) + w_i(k)], \quad i = 3, 4, \cdots, n, \end{aligned} \tag{38}$$

where

$$\begin{aligned} w_1(k) &= [(P^{-1})_{1n}\,\mathrm{Re} - (P^{-1})_{2n}\,\mathrm{Im}]\, f(y_1, \cdots, y_n), \\ w_2(k) &= [(P^{-1})_{2n}\,\mathrm{Re} + (P^{-1})_{1n}\,\mathrm{Im}]\, f(y_1, \cdots, y_n), \end{aligned}$$

in which

$$f(y_1, \cdots, y_n) = e^{j\eta T} \sum_{r=0}^{j+1} b_{r-1}(\xi + j\eta)^{j-r} \cdot \left\{ \exp\left[-(\xi + j\eta) T \operatorname{sat} \frac{d}{T} \left| \sum_{s=1}^{n} y_s(k) \right| \right] - 1 \right\} \cdot \operatorname{sgn} \sum_{s=1}^{n} y_s(k),$$

and $w_i(k)$, $i = 3, 4, \cdots, n$, are defined in (11).

Choosing (12) as a Lyapunov's function for (38), and introducing the polar coordinates defined by (14), we obtain as the sufficient condition for asymptotic stability in the large of $y = 0$ of (38), the following:

$$\gamma_1 \left[e^{2\xi T} \left(\phi_1 \cos \eta T - \phi_2 \sin \eta T + \frac{w_1}{r} \right)^2 - \phi_1^2 \right] + \gamma_2 \left[e^{2\xi T} \left(\phi_1 \sin \eta T + \phi_2 \cos \eta T + \frac{w_2}{r} \right)^2 - \phi_2^2 \right] + \sum_{i=3}^{n} \gamma_i \left[e^{2\alpha_i T} \left(\phi_i + \frac{w_i}{r} \right)^2 - \phi_i^2 \right] < 0$$

$$\text{for all } r \neq 0,\ \phi_i,\ i = 1, \cdots, n, \quad (39)$$

where

$$w_1 = [(P^{-1})_{1n} \operatorname{Re} - (P^{-1})_{2n} \operatorname{Im}] f'(r, \phi),$$
$$w_2 = [(P^{-1})_{2n} \operatorname{Re} + (P^{-1})_{1n} \operatorname{Im}] f'(r, \phi),$$
$$w_1 = \Gamma_i \frac{\exp\left[-\alpha_i T \operatorname{sat} \frac{d}{T} r |\phi| \right] - 1}{\alpha_i |\phi|} \phi, \qquad i = 3, 4, \cdots, n,$$

and

$$f'(r, \phi) = \sum_{r=1}^{j+1} b_{r-1}(\xi + j\eta)^{j-r} e^{j\eta T} \cdot \frac{\exp\left[-(\xi + j\eta) T \operatorname{sat} \frac{d}{T} r |\phi| \right] - 1}{|\phi|} \phi.$$

Unfortunately, the intervals, within which the first two terms of (39) are monotonic functions of r, cannot immediately be determined since it requires solution of transcendental equations. Thus, deduction of explicit sufficient conditions becomes unfeasible. However, in a special case when $\eta T \ll 1$, $f'(r, \phi)$ is reduced to the following:

$$f'(r, \phi) \cong \sum_{r=1}^{j+1} b_{r-1}(\xi + j\eta)^{j-r} \cdot \frac{\exp\left[-\xi T \operatorname{sat} \frac{d}{T} r |\phi| \right] - 1}{|\phi|} \phi$$

and determination of the intervals becomes immediate. The remainder of the procedure of obtaining the sufficient condition follows the same line as that in the case of real simple poles.

B. Multiple Poles

For simplicity of illustration, we assume that the linear plant has one double pole and that the remaining poles are real and simple. Then, the transformation matrix P in (6) should be chosen in such a way that the matrix J is

$$J = \begin{bmatrix} \alpha_2 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \alpha_2 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & \alpha_3 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & \alpha_n \end{bmatrix}.$$

Following the same procedure as that in the previous case we obtain a set of n first-order difference equations as follows:

$$y_1(k+1) = e^{\alpha_2}[y_1(k) + T y_2(d) + w_1(k)]$$
$$y_1(k+1) = e^{\alpha_i}[y_i(k) + w_i(k)], \qquad i = 2, 3, \cdots, n \quad (40)$$

where

$$w_1(k) = \left\{ \sum_{r=0}^{j} b_r \alpha_2^{j-r} \left[(P^{-1})_{1n} - (P^{-1})_{2n} \left(\frac{j-r-1}{\alpha_2} - T \right) \right] \cdot \frac{\exp\left[-\alpha_2 T \operatorname{sat} \frac{d}{T} \left| \sum_{s=1}^{n} y_s(k) \right| \right] - 1}{\alpha_2} - \sum_{r=0}^{j} b_r \alpha_2^{j-r} (P^{-1})_{2n} \cdot \frac{T \operatorname{sat} \frac{d}{T} \left| \sum_{s=1}^{n} y_s(k) \right| \exp\left[-\alpha_2 T \operatorname{sat} \frac{d}{T} \left| \sum_{s=1}^{n} y_s(k) \right| \right]}{\alpha_2} \right\} \cdot \operatorname{sgn} \sum_{s=1}^{n} y_s(k),$$

and $w_i(k)$, $i = 2, 3, \cdots, n$, are defined in (11).

Choosing (12) as a Lyapunov's function and introducing the polar coordinates (14), we obtain the following inequality as the sufficient condition for asymptotic stability in the large:

$$\gamma_1 \left[e^{2\alpha_2 T} \left(\phi_1 + T\phi_2 + \frac{w_i}{r} \right)^2 - \phi_1^2 \right] + \sum_{i=2}^{n} \gamma_i \left[e^{2\alpha_i T} \left(\phi_i + \frac{w_i}{r} \right)^2 - \phi_i^2 \right] < 0$$

$$\text{for all } r \neq 0,\ \phi_i,\ i = 1, 2, \cdots, n, \quad (41)$$

where

$$\frac{w_1}{r} = \sum_{r=0}^{j} b_r \alpha_2^{j-r} \left[(P^{-1})_{1n} - (P^{-1})_{2n} \left(\frac{j-r-1}{\alpha_2} - T \right) \right] \cdot \frac{\exp\left[-\alpha_2 T \operatorname{sat} \frac{d}{T} r \left| \phi \right| \right] - 1}{\alpha_2 r \left| \phi \right|} \phi - \sum_{r=0}^{j} b_r \alpha_2^{j-r} (P^{-1})_{2n} \cdot \frac{T \operatorname{sat} \frac{d}{T} r \left| \phi \right| \exp\left[-\alpha_2 T \operatorname{sat} \frac{d}{T} r \left| \phi \right| \right]}{\alpha_2 r \left| \phi \right|} \phi,$$

$$\frac{w_i}{r} = \Gamma_i \frac{\exp\left[-\alpha_i T \operatorname{sat} \frac{d}{T} r \left| \phi \right| \right] - 1}{\alpha_i r \left| \phi \right|} \phi, \qquad i = 2, 3, \cdots, n.$$

Again note that the intervals, within which the first terms of (41) are a monotonic function of r, cannot be determined immediately for the same reason as before. But, with an additional assumption that $|\alpha_2| T \ll 1$, w_1/r is reduced to

$$\frac{w_1}{r} \cong - \sum_{r=0}^{j} b_r \alpha_2^{j-r} \left[(P^{-1})_{1n} - (P^{-1})_{2n} \left(\frac{j-r-2}{\alpha_2} - T \right) \right] \frac{\operatorname{sat} \frac{d}{T} r \left| \phi \right|}{r \left| \phi \right|} \phi,$$

and determination of such intervals becomes immediate. The remainder of the procedure is the same as that for the previous case.

Acknowledgment

The material in this paper is based on a part of the dissertation submitted by T. T. Kadota in partial satisfaction of the requirements for the Ph.D. degree in Electrical Engineering at the University of California, Berkeley. The authors wish to thank Professors E. I. Jury and S. P. Diliberto for their valuable suggestions and the staff of Electronics Research Laboratory, University of California, for their kind assistance.

Stability and Graphical Analysis of First-Order Pulse-Width-Modulated Sampled-Data Regulator Systems*

E. POLAK†

Summary—**Pulse-width-modulated sampled-data systems are described by nonlinear difference equations which do not lend themselves to an exact analytic treatment.**

This paper presents a graphical technique for the analysis of PWM sampled-data systems with first-order plants. This technique provides a sufficient condition for asymptotic stability in the large, a method for examining the damping properties of the system, a method for computing the step response from any initial condition and, finally, a method for observing and interpreting the effect of varying the system parameters on the step response of the system.

* Received by the PGAC, November 4, 1960; revised manuscript received, April 20, 1961.

† Dept. of Elec. Engrg., University of California, Berkeley, Calif.

I. Introduction

THE behavior of PWMSD systems is described by nonlinear difference equations and, as a rule, cannot be analyzed exactly. Amongst the methods used in the approximate analysis of these systems we find: describing function techniques [2], linearization [3], and techniques based on Lyapunov's second method [4, 5]. However, none of these methods is easy to apply and the results they yield are usually rather limited.

Both describing functions and linearization techniques can be used in examining a system for local stability as well as for predicting its performance for small deviations from the equilibrium point. Their usefulness lies mainly in the analysis of systems with sampling rates which are high compared to the dominant time constant of the system. However, these techniques become rather unrealiable when applied to a system with slow sampling and with a broad range of operation.

Techniques based on Lyapunov's second method have been used to test for system stability in the large [5] as well as in the choosing of system parameters in system synthesis [4]. They can also be used to obtain a qualitative measure of the damping rate of the transient response. However, at the present state of the art, the accuracy obtained is very poor, usually giving very large margins of safety on the parameter values required to ensure asymptotic stability in the large.

The only exception to this general rule of intractability is found in first-order systems, for which a very simple graphical method of analysis is presented in this paper. It was suggested by a technique used by Madwed [6] to solve, numerically, nonlinear differential equations and it enables one to construct a step response very quickly, examine the system for asymptotic stability in the large, estimate the rate of damping of the transient response, and provide a simple means for observing the effect of changing various system parameters.

One particular system will be analyzed in full in the body of the paper and the equations necessary for the various graphical constructions for two other systems will be derived in the Appendix. It will be observed that the method is quite general and applies to any one of the system types examined with equal ease.

II. Statement of the Problem

For the regulator-type system described in Section III, it is required to develop a simple graphical technique which will enable one to compute

1) The open-loop step response
2) The closed-loop step response
3) Whether sufficient conditions for asymptotic stability in the large are satisfied
4) An approximation to the damping coefficient in the unsaturated region of operation.

III. Description of the System

The flow diagram of the system to be considered is shown in Fig. 1. The system consists of a plant, an error detector-input element, an error amplifier, and a modulator. The block diagram of the system using Laplace transform notation is given in Fig. 2. The symbols used in the block diagram will now be defined:

$\theta(t)$ is the value of the system output at time t.
$\theta_e(t)$ is the value of the error at time t.
θ_d is the desired value, a constant.
k is the error amplifier gain.
T is the plant time constant.
$X(t)$ is the modulator output at time t.

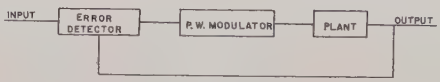


Fig. 1—Flow diagram of system.

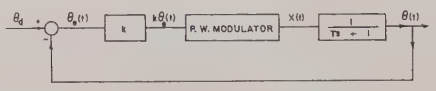


Fig. 2—Block diagram of system.

It will be observed that a unit gain has been assigned to the plant since it is possible to incorporate this gain into the modulator description in a rather obvious manner, as will be noticed later.

IV. Description of the Pulse-Width Modulator

The output of the modulator considered is described by the following modulation law:

$$X(t) = \begin{cases} (\operatorname{sgn}\theta_{en})\cdot\theta_m & \text{for } nT_s \le t \le nT_s + \alpha_n T_s \\ 0 & \text{for } nT_s + \alpha_n T_s \le t \le (n+1)T_s \end{cases} \quad (1)$$

where

θ_m = the modulator pulse height (corrected for plant gain)
$\theta_{en} = \theta_e(nT_s)$
T_s = the sampling period
$\alpha_n T_s$ = the pulse width for the $(n+1)$th sampling period: $nT_s \le t \le (n+1)T_s$.

α_n is defined as follows:

$$\alpha_n = \operatorname{Sat} \left| (k\theta_{en}) \right| \qquad \text{for } nT_s \le t \le (n+1)T_s \quad (2)$$

where

$$\operatorname{Sat} x = \begin{cases} x & \text{for } 0 \le x \le 1 \\ 1 & \text{for } x > 1 \\ 0 & \text{for } x < 0 \end{cases}.$$

A typical input-output relation for the modulator is given in Fig. 3.

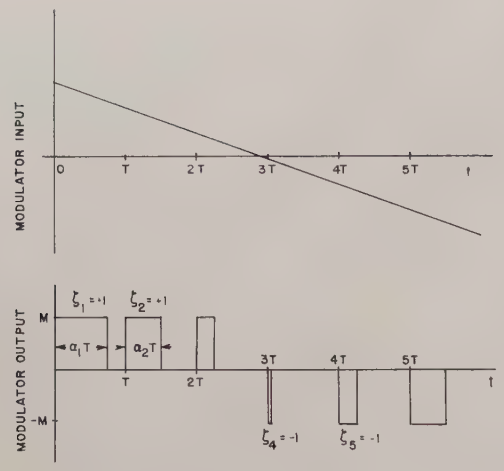


Fig. 3—Pulse-width-modulator input-output relations.

V. Graphical Computation of Open-Loop Step Response

A. Derivation of Difference Equation

Under open-loop conditions, the feedback line is broken and the output of the error amplifier assumes the constant value $k\theta_d$.

Defining

$$\alpha_0 = \text{Sat} \mid k\theta_d \mid , \qquad 0 \leq \alpha_0 \leq 1$$

and

$$\theta_n = \theta(nT_s) \tag{3}$$

we obtain the difference equation in the following manner. For $nT_s \leq t \leq (\alpha_0 + n)T_s$ the differential equation describing the system is given by

$$T\dot{\theta} + \theta = [\text{sgn } \theta_d]\theta_m \cdot H(t - nT_s). \tag{4}$$

The solution of this equation with the initial condition that at $t = nT_s$ $\theta(t) = \theta_n$ is given by

$$\theta(t') = \theta_n e^{-t'/T} + [\text{sgn } \theta_d] \cdot \theta_m (1 - e^{-t'/T}) \tag{5}$$

where

$$t' = t - nT_s. \tag{6}$$

Evaluating $\theta(t')$ for $t' = \alpha_0 T_s$, we get

$$\theta(\alpha_0 T_s) = \theta_n e^{-\alpha_0 T_s/T} + [\text{sgn } \theta_d] \cdot \theta_m (1 - e^{-\alpha_0 T_s/T}). \tag{7}$$

Now for the rest of the $(n+1)$th sampling period, the differential equation becomes, *i.e.*, for

$$nT_s + \alpha_0 T_s \leq t \leq (n+1)T_s, \qquad T\dot{\theta} + \theta = 0. \tag{8}$$

The solution of this equation with $\theta(\alpha_0 T_s)$ as the initial value is found to be

$$\theta(t') = \theta_n e^{-t'/T} + [\text{sgn } \theta_d] \cdot \theta_m e^{-t'/T}(e^{\alpha_0 T_s/T} - 1). \tag{9}$$

Finally, putting $t' = T_s$, *i.e.*, $t = (n+1)T_s$, in (9) we obtain the difference equation describing the open-loop step response as

$$\theta_{n+1} = \theta_n e^{-T_s/T} + [\text{sgn } \theta_d]\theta_m e^{-T_s/T}(e^{\alpha_0 T_s/T} - 1). \tag{10}$$

The above is a simple linear difference equation and can be solved either analytically or graphically with about equal ease.

B. Graphical Solution of Difference Equation

One begins by drawing (10) in the (θ_n, θ_{n+1}) plane. This results in a straight-line locus which will be referred to as the OPEN-LOOP CHARACTERISTIC (OLC). Since at steady state $\theta_{n+1} = \theta_n$, the intersection of the OLC with the line $\theta_{n+1} = \theta_n$ gives the steady-state solution. This will be referred to as θ_s. The step-by-step values of the response from an initial condition θ_0 are found as follows. Given θ_0 on the θ_n axis, θ_1 is found on the OLC, vertically above θ_0; it is then transferred back to the θ_n axis by reflecting it about the $\theta_{n+1} = \theta_n$ line. The process is then repeated to obtain θ_2, etc. If all the construction lines are drawn in, the plot resembles a staircase as shown in Fig. 4. It will be observed that the graphical construction is extremely simple and that it gives a rather interesting representation of the step response differing markedly from the usual amplitude-time-type plot.

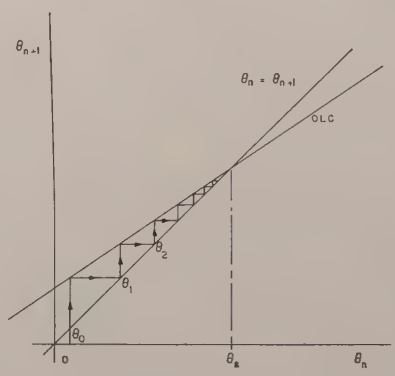


Fig. 4—Typical open-loop step response.

VI. Graphical Computation of Closed-Loop Step Response

A. Derivation of Difference Equation

Under closed-loop conditions, the modulation coefficient is no longer constant, but is a function of the error. As a result of this the describing difference equation becomes nonlinear as will presently be shown.

Substituting α_n for α_0 and θ_{en} for θ_d in (10), as well as expanding for α_n according to (2), we obtain the difference equation for the closed-loop step response. This is given by

$$\theta_{n+1} = \theta_n e^{-T_s/T} + [\text{sgn } (\theta_d - \theta_n)] \cdot \theta_m e^{-T_s/T}(e^{|k(\theta_d - \theta_n)|T_s/T} - 1) \quad \text{for } \mid k\theta_{en} \mid \leq 1 \tag{11}$$

and

$$\theta_{n+1} = \theta_n e^{-T_s/T} + [\text{sgn } (\theta_d - \theta_n)] \cdot \theta_m (1 - e^{-T_s/T}) \quad \text{for } \mid k\theta_{en} \mid > 1. \tag{12}$$

These equations are nonlinear, and cannot be solved explicitly for either the steady-state value θ_s (*i.e.*, $\theta_n=\theta_{n+1}=\theta_s$) or for the step-by-step values of the step response. They can be solved, however, by either numerical or graphical techniques.

B. Graphical Solution of Difference Equation

As in the linear case, one first plots (11) and (12) in the (θ_n, θ_{n+1}) plane (see Fig. 5). This locus defines the closed-loop step response from any initial value θ_0 and will be called the CLOSED-LOOP CHARACTERISTIC (CLC). One also draws the line $\theta_{n+1}=\theta_n$.

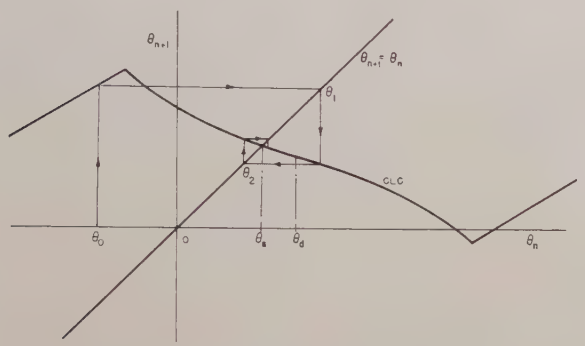


Fig. 5—Typical closed-loop step response.

The intersection of the CLC with the line $\theta_{n+1}=\theta_n$ gives the value of the equilibrium point θ_s which may now be either stable or unstable, depending on the system parameters.

The step-by-step response from an initial value θ_0 is obtained in a manner analogous to the one used for the open-loop case. One reads off the value of θ_1 on the CLC, vertically above θ_0, and then one transfers this value back onto the θ_n axis by means of a reflection about the $\theta_{n+1}=\theta_n$ line. The process is then repeated to find θ_2, etc. The resulting graph in the (θ_n, θ_{n+1}) plane is either staircase-like, or else resembles a spiderweb spiralling either in or out, depending on whether the system is stable or not. It may also be a combination of the two above-mentioned cases. A plot of a typical step response is shown in Fig. 5.

VII. Sufficient Condition for Asymptotic Stability in the Large

Theorem

A sufficient condition for asymptotic stability in the large is that there exist a number $\epsilon>0$ such that

1) The CLC lies below the straight line K_1 and above the straight line K_2 for $\theta_n<\theta_s$, and
2) The CLC lies above the straight line K_1 and below the straight line K_2 for $\theta>\theta_s$.

where

K_1 is defined by $\theta_{n+1} = -(1-\epsilon)\theta_n + (2-\epsilon)\theta_s$ (13)

and

K_2 is defined by $\theta_{n+1} = (1-\epsilon)\theta_n + \epsilon\theta_s$. (14)

Proof:

Referring to Fig. 6, consider any $\theta_n \neq \theta_s$, $|\theta_s| < \infty$ $|\theta_0| < \infty$, then, if the conditions 1) and 2) above are satisfied, the following inequality holds:

$$|\theta_{n+1} - \theta_s| \leq (1-\epsilon)|\theta_n - \theta_s|$$

and, hence,

$$\begin{aligned} |\theta_{n+1} - \theta_s| &\leq (1-\epsilon)^2 |\theta_{n-1} - \theta_s| \\ &\vdots \\ &\leq (1-\epsilon)^{n+1} |\theta_0 - \theta_s| \end{aligned} \quad (15)$$

but $|\theta_0| < \infty$ and $|\theta_s| < \infty$ by assumption and $(1-\epsilon)^{n+1} \to 0$ as $n \to \infty$. Hence

$$|\theta_{n+1} - \theta_s| \to 0 \quad \text{as} \quad n \to \infty$$

and

$$\theta_{n+1} \to \theta_s \quad \text{as} \quad n \to \infty.$$

This proves that the system is asymptotically stable in the large, Q.E.D.

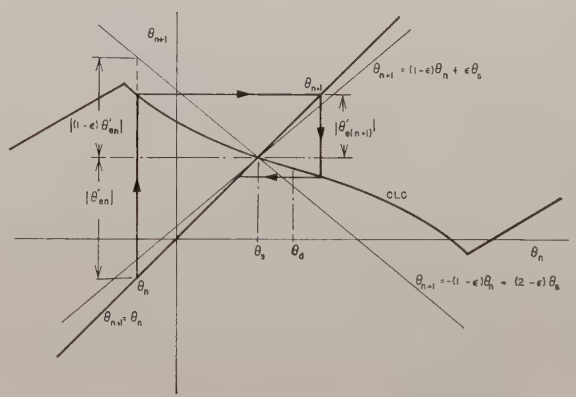


Fig. 6—Test for asymptotic stability in the large

That the condition given in the theorem is only a sufficient condition for asymptotic stability in the large can be simply demonstrated by constructing a CLC which does not satisfy the condition, but which nevertheless does describe a system which is asymptotically stable in the large. Such a CLC is shown in Fig. 7 and the stability of the system is self-evident.

Limited experimentation with the three systems considered in this paper indicates that the gain of a system which just satisfies the condition of stability can be increased no more than 20 per cent before the system actually ceases to be asymptotically stable in the large. This indicates that the condition is not too conservative for practical purposes.

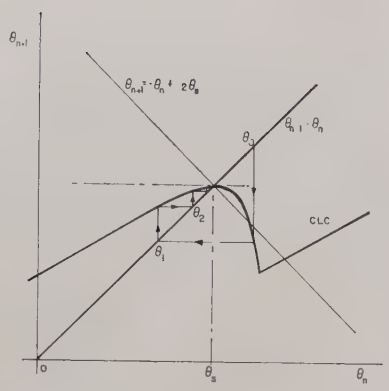


Fig. 7—Example of system asymptotically stable in the large and not satisfying sufficient condition for stability.

VIII. Estimation of Damping Properties of the Step Response in the Unsaturated Region of Operation

A. Frequency of Damped Oscillations

From the construction of an oscillatory response in the (θ_n, θ_{n+1}) plane, it will be observed that for oscillations to occur the CLC must have regions of negative slope on both sides of θ_s. If the slope of the CLC is negative in the entire unsaturated region, then the frequency of the oscillation will be $1/2T_s$. It may be interesting to observe at this point that if a first-order *linear* difference equation has an oscillatory solution then its frequency is also half the sampling frequency [7].

If the slope of the CLC in the unsaturated region is positive for some θ_n and negative for other θ_n then the nature of the response may, although it need not, become more complex, *e.g.*, it may go through a few oscillations and then continue as an overdamped response.

B. Estimation of Rate of Damping

Consider the slope of the CLC for $k|\theta_n - \theta_d| \leq 1$.

$$\frac{d\theta_{n+1}}{d\theta_n} = e^{-T_s/T}\left(1 - \frac{kT_s}{T}\theta_m e^{k|\theta_{en}|T_s/T}\right). \tag{16}$$

It is obvious by inspection that depending on the values of the parameters, the slope may remain positive within the whole region under consideration, positive for small values of θ_{en} and negative for large values of θ_{en}, and, finally, it may remain negative throughout the whole unsaturated region.

Case I—Slope Does Not Change Sign in Unsaturated Region: The procedure for obtaining the damping rate estimate is as follows. Draw a straight line approximation to the CLC in the unsaturated region as shown in Fig. 8. This line makes an angle ϕ with the θ_n axis. It is now possible to write down the linearized difference equation for the system for $k|\theta_n - \theta_s| \leq 1$. Putting

$$|\theta_n - \theta_s| = |\theta_{en}'| \tag{17}$$

it is found from the geometry of Fig. 7 that

$$|\theta'_{e(n+1)}| = |\theta_{en}'| \cdot |\tan\phi|. \tag{18}$$

The solution of this difference equation with θ_{en}' as the initial value from which the computation is started in the unsaturated region is given by

$$|\theta_{en}'| = |\theta_{e0}'| \cdot \exp\left[nT_s \frac{\ln|(\tan\phi)|}{T_s}\right] \tag{19}$$

where $\ln|(\tan\phi)|/T_s$ may be interpreted as a damping coefficient. It will be observed that the smaller the value of $|\tan\phi|$ the faster will the transient die out. It will also be noticed that in the form in which the equations are written they are equally applicable to systems with either positive slope CLC's or negative slope CLC's in the unsaturated region.

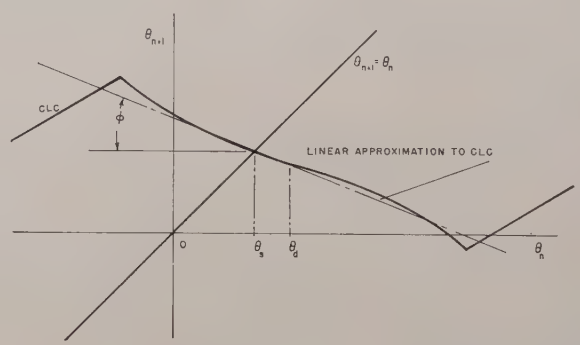


Fig. 8—Graphical estimation of damping.

Case II—The Slope of the CLC Changes Sign in the Unsaturated Region: Now the unsaturated region has to be broken up into two subregions, one in which the response is oscillatory and the other one in which the response is overdamped. After approximating the CLC in each of these regions by a straight line, one proceeds as before to find the damping coefficient of each region.

IX. Evaluation of Critical Gain Value k_c

In the design of PWMSD systems, the sampling frequency is usually made as high as possible and its value is chosen on the basis of limitations imposed by hardware. The effective pulse height is chosen on the basis of speed of response requirements in the saturated region as well as on the basis of range requirements. Thus the only free parameter left for adjusting the shape of the CLC in the unsaturated region is the error amplifier gain k.

The procedure for finding the critical gain value k_c, *i.e.*, the maximum value of k for which the sufficient condition for asymptotic stability in the large is satis-

fied, is one of successive approximation. However, as it will be observed, this procedure is a very fast one and quite easy to carry out.

The graphical construction of the new CLC may be carried out easily by means of the following graphical method. Draw a number of equispaced lines parallel to the saturated portions of the CLC and intersecting the unsaturated region of the CLC at the points θ_{n1} as shown in Fig. 9. One now obtains new values for θ_n, θ_{n2}, by using the formula

$$k_{c1}(\theta_d - \theta_{n1}) = k_{c2}(\theta_d - \theta_{n2}) \tag{20}$$

and then finds the new corresponding points θ_{n+1} at the intersection of the lines $\theta_n = \theta_{n2}$ with the lines through θ_{n1} as is again shown in Fig. 9. It will be noted that the same graphical procedure could have been used in shaping the CLC for a satisfactory transient response.

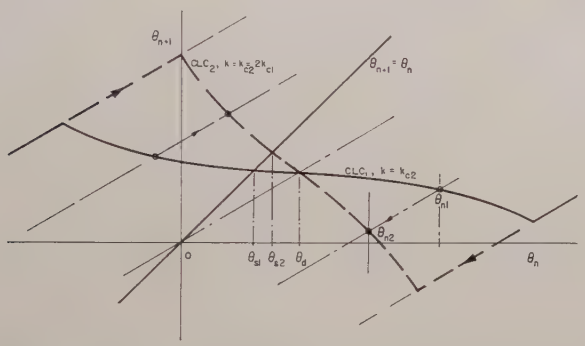


Fig. 9—Graphical construction of the CLC from another one.

X. Conclusions

The main merit of the graphical technique described is that the CLC defines the dynamic behavior of the closed-loop system completely over any range of initial conditions. Thus, once the CLC and the lines K_1, K_2 are drawn in the (θ_n, θ_{n+1}) plane, it is apparent at a glance whether the system does or does not satisfy the sufficient condition for asymptotic stability in the large, whether damped oscillations can or cannot take place and, if they can occur, for which initial values of θ_n. It is equally easy to determine the approximate rate at which the transient decays in a stable system.

As a tool of synthesis, the graphical technique is mainly suitable for examining the effects of varying the error amplifier gain or the effective modulator pulse height on the performance of the system. The effect of varying the sampling frequency can also be studied graphically, but the procedure becomes more cumbersome than in the case of gain or pulse height.

In conclusion it should be noted that the methods presented in this paper do not carry over to higher-order systems. However, should it be possible to approximate a higher-order system by its dominant time constant, then they may be useful for an approximate analysis.

Appendix

Derivation of Equations for OCL and CLC for Two Other Systems

In this Appendix will be derived the equations for the OLC and the CLC for two systems using modulators different from the one described in the body of this paper. It will be observed that these equations are quite similar to the ones obtained for the first system and can therefore be solved by the same graphical technique.

I. Systems With Modulators of Type II

A. Description of Modulator

This particular type of modulator is very frequently used in hydraulic servos, because it is easier to build than the one previously described and because it offers advantages if the plant has stiction. Its action consists of modulating the symmetry of a square wave and its output is defined by the following equations:

$$X(t) = \begin{cases} \theta_m & \text{for } nT_s \le t \le nT_s + \beta_n T_s \\ -\theta_m & \text{for } nT_s + \beta_n T_s \le t \le (n+1)T_s \end{cases} \tag{21}$$

where

$$\beta_n = \text{Sat } (f_1(\theta_d) + k\theta_{en}) \tag{22}$$

and

$$\text{Sat } (x) = \begin{cases} x & \text{for } 0 \le x \le 1 \\ 1 & \text{for } x > 1 \\ 0 & \text{for } x < 0 \end{cases}.$$

$f_1(\theta_d)$ is a bias function serving the purpose of eliminating the steady-state error and will be investigated later.

B. Equation for the OLC

For the interval $0 \le t' \le T_s$, *i.e.*, $nT_s \le t \le (n+1)T_s$, the differential equation describing the system open-loop step response is given by

$$T\dot{\theta} + \theta = \theta_m H(t') - 2H(t' - \beta_0 T_s) \tag{23}$$

where

$$\beta_0 = \text{Sat } [f_1(\theta_d) + k\theta_d]. \tag{24}$$

The solution of this equation with θ_n as the initial value is given by

$$\theta(t') = \theta_n e^{-t/T} - \theta_m[1 - e^{-t'/T}(2e^{\beta_0 t'/T} - 1)]. \tag{25}$$

Evaluating $\theta(t')$ for $t' = T_s$, we obtain the equation for the OLC:

$$\theta_{n+1} = \theta_n e^{-T_s/T} - \theta_m[1 - e^{-T_s/T}(2e^{\beta_0 T_s/T} - 1)]. \tag{26}$$

As in the previous case, this is a linear first-order difference equation.

C. Equation for the CLC

By substituting β_n for β_0 in (26) and expanding for β_n according to (22) the difference equations for the closed-loop step response in the respective regions of operation are:

for

$$0 \leq f_1(\theta_d) + k\theta_{en} \leq 1,$$

$$\theta_{n+1} = \theta_n e^{-T_s/T} - \theta_m\left[1 - e^{-T_s/T}(2e^{f_1(\theta_d)T_s/T} \cdot e^{k\theta_{en}T_s/T} - 1)\right]; \quad (27)$$

for

$$f_1(\theta_d) + k\theta_{en} < 0, \quad \theta_{n+1} = \theta_n e^{-T_s/T} - \theta_m(1 - e^{-T_s/T}); \quad (28)$$

and, finally, for

$$f_1(\theta_d) + k\theta_{en} > 1, \quad \theta_{n+1} = \theta_n e^{-T_s/T} + \theta_m(1 - e^{-T_s/T}). \quad (29)$$

The CLC may now be plotted from these equations as before.

D. Equation for Bias Function $f_1(\theta_d)$

The purpose of the bias function $f_1(\theta_d)$ used in this type of modulator is to eliminate steady-state error. It is thus determined by the condition that for all θ_d, the steady-state solution of (27) $\theta_s = \theta_d$. Combining this condition with (27) we get

$$\theta_d = \theta_d e^{-T_s/T} - \theta_m\left[1 - e^{-T_s/T}(2e^{f_1(\theta_d)T_s/T} - 1)\right]. \quad (30)$$

Solving the above equation for $f_1(\theta_d)$ we get

$$f_1(\theta_d) = \frac{T_s}{T} \ln\left[\frac{\theta_d}{2\theta_m}(e^{T_s/T} - 1) + \tfrac{1}{2}(e^{T_s/T} + 1)\right]. \quad (31)$$

II. Systems with Modulators of Type III

A. Description of Modulator

This modulator is a special case of the type II modulator and is used in cases where pulses of one polarity only can be generated, as in heating and flow control systems. Its output is defined as follows:

$$X(t) = \begin{cases} \theta_m & \text{for } nT_s \leq t \leq \gamma_n T_s + nT_s \\ 0 & \text{for } nT_s + \gamma_n T_s \leq t \leq (n+1)T_s \end{cases} \quad (32)$$

where

$$\gamma_n = \text{Sat}\left[f_2(\theta_d) + k\theta_{en}\right] \quad (33)$$

and $f_2(\theta_d)$ is a bias function.

B. Equation for the OLC

For the interval $0 \leq t' \leq T_s$, the differential equation describing the system open step response is given by

$$T\dot{\theta} + \theta = \theta_m[H(t') - H(t' - \gamma_0 T_s)] \quad (34)$$

where

$$\gamma_0 = \text{Sat}\left[f_2(\theta_d) + k\theta_d\right]. \quad (35)$$

The solution of this equation is given by

$$\theta(t') = \theta_n e^{-t'/T} + \theta_m e^{-t'/T}(e^{\gamma_0 t'/T} - 1). \quad (36)$$

Now evaluating $\theta(t')$ for $t' = T_s$, we obtain the equation for the OLC:

$$\theta_{n+1} = \theta_n e^{-T_s/T} + \theta_m e^{-T_s/T}(e^{\gamma_0 T_s/T} - 1). \quad (37)$$

C. Equation for the CLC

As before, we obtain the equations for the CLC from the equation for the OLC by substituting γ_n for γ_0 and expanding for γ_n according to (33). We thus find for

$$0 \leq f_2(\theta_d) + k\theta_{en} \leq 1,$$

$$\theta_{n+1} = \theta_n e^{-T_s/T} + \theta_m e^{-T_s/T}(e^{f_2(\theta_d)T_s/T} \cdot e^{k(\theta_d - \theta_n)T_s/T} - 1); \quad (38)$$

for

$$f_2(\theta_d) + k\theta_{en} < 0, \qquad \theta_{n+1} = \theta_n e^{-T_s/T}; \quad (39)$$

and for

$$f_2(\theta_d) + k\theta_{en} > 1, \quad \theta_{n+1} = \theta_n e^{-T_s/T} + \theta_m(1 - e^{-T_s/T}). \quad (40)$$

Again, the graphical solution proceeds as before.

D. Equation for the Bias Function $f_2(\theta_d)$

Again, as before, $f_2(\theta_d)$ is defined so that there should be no steady-state error for all θ_d and is thus found by putting $\theta_{n+1} = \theta_n = \theta_d$ in (18) as follows:

$$\theta_d = \theta_d e^{-T_s/T} + \theta_m e^{-T_s/T}(e^{f_2(\theta_d)T_s/T} - 1). \quad (41)$$

Solving this equation for $f_2(\theta_d)$, we get

$$f_2(\theta_d) = \frac{T_s}{T} \ln\left[\frac{\theta_d}{\theta_m}(e^{T_s/T} - 1) + 1\right]. \quad (42)$$

Acknowledgment

The author wishes to express his appreciation to Prof. C. A. Desoer and Prof. A. M. Hopkin of the University of California, Berkeley, for their helpful suggestions and comments.

References

[1] E. Polak, "Stability Criteria for Pulse Width Modulated Sampled Data Systems," M.S. thesis, University of California, Berkeley; 1959.
[2] J. G. Chubbuck, "Are high performance and low cost compatible in hydraulic servos?," *Control Engrg.*, vol. 4, pp. 98–103; March, 1957.
[3] R. E. Andeen, "Analysis of Pulse Duration Sampled-Data Systems with Linear Elements," Ph.D. dissertation, Northwestern University, Evanston, Ill.; 1958.
[4] W. L. Nelson, "Pulse Width Control of Sampled Data Systems," Ph.D. dissertation, Columbia University, New York, N. Y.; 1959.
[5] T. T. Kadota, "Analysis of Nonlinear Sampled-Data Systems with Pulse Width Modulators," Ph.D. dissertation, University of California, Berkeley; 1960.
[6] A. Madwed, "Number Series Methods of Solving Linear and Nonlinear Differential Equations," Ph.D. dissertation, Mass. Inst. Tech., Cambridge; 1950.
[7] Charles Jordan, "Calculus of Finite Differences," Chelsea Publishing Co., New York, N. Y., 2nd ed., pp. 552–554; 1950.
[8] H. Levy and F. Lessman, "Finite Difference Equations," The Macmillan Co., New York, N. Y., ch. 5; 1961.

Analysis of Pulse-Width-Modulated Control Systems*

F. R. DELFELD†, AND G. J. MURPHY‡, MEMBER, IRE

Summary—Previous work on the analysis and design of pulse-width-modulated control systems is reviewed, and the limitations of some of the earlier contributions are discussed. A mathematical development of an orderly and relatively simple method for the exact determination of the response of closed-loop pulse-width-modulated control systems to arbitrary input is then presented. Through the use of difference equations and the separation of linear and nonlinear terms, the output at the sampling instants is expressed as a function of the sampled error, and z-transform theory is then employed to obtain an exact solution for the error at the sampling instants.

A technique for studying the stability and other performance characteristics of pulse-width-modulated feedback systems is next presented. The exact method of analysis developed earlier and a modified describing-function technique are utilized together to investigate stability without overlooking pulse-width saturation.

Three illustrative examples are also presented to demonstrate the relative simplicity of the methods described in the paper as well as the accuracy of the results obtained by the use of these methods.

Introduction

A PULSE-WIDTH-MODULATED control system is a sampled-data system in which the duration of each pulse is made proportional to the absolute value of the sampled signal at the sampling instants. All pulses are rectangular and of the same amplitude, and the sign of a given pulse is the same as that of sampled signal at the corresponding sampling instants. The pulse-width modulator which produces this pulse sequence is nonlinear.

The exact difference equations for the pulse-width-modulated control system relating the controlled variable $c(t)$ to the system error $e(t)$ at the sampling instants were considered by Tsypkin[1] and Nease.[2] Tsypkin[3] provided a general method for the exact analysis of pulse-width-modulated systems by direct superposition of appropriately weighted step-function responses. Da-Chuan[4] developed a criterion for the existence of self-oscillations of a pulse-width modulated control system with a period equal to twice the sampling period.

For sufficiently small inputs, the pulse-width modulator output may be approximated with a sequence of impulse functions of equal area.[1,2,5,6] With this approximation, the pulse-width-modulated control system is equivalent to the conventional linear sampled-data system and may be studied by use of conventional sampled-data theory. Nelson[7] applied Lyapunov's second method to the design of pulse-width-modulated systems using this approximation. However, Nelson's approach is limited by the assumption that the pulse-width is so small in every sampling interval that the actual pulses can be replaced by impulse functions as in the earlier work.[1,2,5,6]

The purpose of this paper is to present both an orderly and relatively simple method for the exact determination of the response of closed-loop pulse-width-modulated systems to arbitrary input and a design technique which is not limited by an assumption of small pulse-widths. Through use of the system difference equations and separation of linear and nonlinear terms, the output at the sampling instants is expressed as a function of the sampled error. Application of z-transform theory leads to an exact solution for the error at the sampling instants. Infinite series expansions and matrix equations similar to those obtained by Kinnen and Tou[8] are employed in this development.

A technique for studying the stability and performance of pulse-width-modulated systems is presented through use of the exact response method and a modified describing-function method. In the modified describing-function method, describing-function techniques are applied to an equivalent nonlinear sampled-data system which is obtained from the exact difference equations. The application of the exact-response method and the modified describing function in an illustrative example reveals the presence of unstable and stable limit cycles and demonstrates the limitations of the small pulse-width approximation.

* Received by the PGAC, October 21, 1960; revised manuscript received, April 17, 1961. This paper is an outgrowth of a dissertation submitted by F. R. Delfeld to the Graduate School of Northwestern University in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

† A-C Spark Plug Div., General Motors Corp., Milwaukee, Wis.

‡ Elec. Engrg. Dept., Northwestern University, Evanston, Ill.

[1] E. P. Popow, "Dynamik Automatischer Regelsysteme," Akademie-Verlag, Berlin, Ger., pp. 407–409; 1958.

[2] R. F. Nease, "Analysis and Design of Non-linear Sampled-Data Control Systems," Mass. Inst. Tech., Cambridge, WADC Tech. Note 57-162; 1957.

[3] Ja. S. Tsypkin, "The calculation of the process in nonlinear impulse control systems" (in Russian), *Avtomat. i Telemekh.*, vol. 12, pp. 500–512; June, 1956.

[4] S. Da-Chuan, "On the possibility of certain types of oscillations in sampled-data control systems," *Avtomat. i Telemekh.*, vol. 20, pp. 85–89; January, 1959.

[5] W. W. Solodownikow, "Grundlagen der Selbsttätigen Regelung," R. Oldenbourg Verlag, Munich, Ger., pt. 1, ch. 21; 1958.

[6] R. E. Andeen, "Analysis of pulse duration sampled-data systems with linear elements," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-5, pp. 306–313; September, 1960.

[7] W. L. Nelson, "Pulse-width relay control in sampling systems," *Trans. ASME*, vol. 83, pp. 65–76; March, 1961.

[8] E. Kinnen and J. Tou, "Analysis of nonlinear sampled-data systems—part I," *Trans. AIEE (Applications and Industry*, no. 46), pp. 386–394; January, 1960.

System Response Method

A block diagram for a pulse-width-modulated control system is shown in Fig. 1. The pulse-width-modulator output $m(t)$ is defined by the equation

$$m(t) = \begin{cases} A \text{ Sgn } e(nT), & nT \leq t < nT + \alpha \left| e(nT) \right| \quad ; \quad \left| e(nT) \right| < T/\alpha \\ 0 \quad , & nT + \alpha \left| e(nT) \right| \leq t < (n+1)T; \quad \left| e(nT) \right| < T/\alpha \\ A \text{ Sgn } e(nT), & nT \leq t < (n+1)T \quad ; \quad \left| e(nT) \right| > T/\alpha, \end{cases} \tag{1}$$

where $e(nT)$ is the system error at $t=nT$, $n=0, 1, 2, 3, \cdots$; T is the constant sampling period; and A and α are constants determined by the characteristics of the pulse-width modulator.

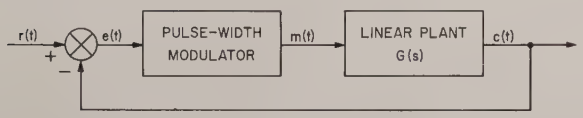


Fig. 1—Model of the pulse-width-modulated control system.

The linear plant is represented by the transfer function $G(s)$, defined by

$$G(s) = \frac{P(s)}{Q(s)}, \tag{2}$$

where $P(s)$ and $Q(s)$ are polynomials in s and the degree of $P(s)$ is at least 1 less than that of $Q(s)$, and $P(s)$ and $Q(s)$ have no common factors. For the case of no repeated zeros of $Q(s)$, the transfer function $G(s)$ may be expressed in partial fraction form as

$$G(s) = \sum_{i=1}^{k} \frac{A_i}{s + a_i}, \tag{3}$$

where

$$A_i = \frac{P(-a_i)}{Q'(-a_i)} \tag{4}$$

and

$$Q'(-a_i) = \left[\frac{d}{ds} Q(s) \right]_{s=-a_i}. \tag{5}$$

The system output at the sampling instants is[2]

$$c(nT) = \sum_{i=1}^{k} c_i(nT), \tag{6}$$

where

$$c_i[(n+1)T] - e^{-a_i T} c_i(nT) = A\alpha A_i e^{-a_i T} \left[\frac{e^{a_i \alpha \left| e_s(nT) \right|} - 1}{\alpha a_i} \right] \text{Sgn } e(nT) \tag{7}$$

and

$$e_s(nT) = \begin{cases} e(nT), & \left| e(nT) \right| \leq T/\alpha \\ \dfrac{T}{\alpha} \text{ Sgn } e(nT), & \left| e(nT) \right| \geq T/\alpha \end{cases}. \tag{8}$$

The difference equations (6) and (7) may be expressed in the form of a vector difference equation,

$$\mathbf{x}(n+1)T = \mathbf{G}(T)\mathbf{x}(nT) + A\alpha \text{ Sgn } e(nT)\mathbf{G}(T - \tau_n)\mathbf{h}(\tau_n), \tag{9}$$

where $\mathbf{x}$ is a k vector with elements x_i, and

$$\mathbf{G}(T) = \begin{bmatrix} e^{-a_1 T} & & & 0 \\ & e^{-a_2 T} & & \\ & & \ddots & \\ 0 & & & e^{-a_k T} \end{bmatrix}, \tag{10}$$

and

$$\mathbf{h}(\tau_n) = \begin{bmatrix} \dfrac{1}{\alpha a_1}(1 - e^{-a_1 \tau_n}) \\ \dfrac{1}{\alpha a_2}(1 - e^{-a_2 \tau_n}) \\ \vdots \\ \dfrac{1}{\alpha a_k}(1 - e^{-a_k \tau_n}) \end{bmatrix}. \tag{11}$$

The state vector $\mathbf{x}$ is related to the system output with

$$c_i(nT) = A_i x_i(nT), \tag{12}$$

$$c(nT) = \sum_{i=0}^{k} A_i x_i(nT). \tag{13}$$

The pulse-width following the nth sampling instant is

$$\tau_n = \alpha \left| e_s(nT) \right|. \tag{14}$$

The nonlinear terms may be isolated by introduction of auxiliary variables $u_i(nT)$ defined by

$$e(nT) + u_i(nT) = \left[\frac{e^{a_i \alpha \left| e_s(nT) \right|} - 1}{\alpha a_i} \right] \text{Sgn } e(nT) \tag{15a}$$

$$= \left[\frac{e^{a_i \tau_n} - 1}{\alpha a_i} \right] \text{Sgn } e(nT). \tag{15b}$$

Use of (15) in (9) gives

$$\mathbf{x}[(n+1)T] = \mathbf{G}(T)\mathbf{x}(nT) + A\alpha\mathbf{G}(T)\mathbf{I}_c e(nT) + A\alpha\mathbf{G}(T)\mathbf{u}(nT), \tag{16}$$

where I_c is a k vector with all elements equal to unity. Eq. (16) can be linearized by the small-signal approximation,

$$\left[\frac{e^{a_i\alpha|e_s(nT)|}-1}{\alpha a_i}\right]\operatorname{Sgn} e(nT) \cong e(nT).$$

With this approximation the auxiliary vector $\boldsymbol{u}$ is reduced to zero and the system may be studied by the use of the conventional methods of sampled-data theory.

However, analysis and design of pulse-width-modulated systems on the basis of this approximation are severely limited because the essential nonlinearity of the system is being overlooked.

Eq. (16) may be expressed in terms of z transforms as

$$z\boldsymbol{X}(z) = \boldsymbol{G}(T)\boldsymbol{X}(z) + A\alpha\boldsymbol{G}(T)\boldsymbol{I}_cE(z) + A\alpha\boldsymbol{G}(T)\boldsymbol{U}(z) \tag{17}$$

provided that $\mathbf{x}(0)=0$. Similarly, (13) can be written as

$$C(z) = \boldsymbol{A}\boldsymbol{X}(z) \tag{18}$$

where

$$\boldsymbol{A} = \begin{bmatrix} A_1 & A_2 & A_3 \cdots & A_k \\ 0 & 0 & 0 & 0 \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{19}$$

Combination of (17) and (18) gives

$$C(z) = \boldsymbol{A}[z\boldsymbol{I} - \boldsymbol{G}(T)]^{-1} \cdot \{A\alpha E(z)\boldsymbol{G}(T)\boldsymbol{I}_c + A\alpha\boldsymbol{G}(T)\boldsymbol{U}(z)\}, \tag{20}$$

where

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & 0 \cdots 0 \\ 0 & 1 & 0 \cdots 0 \\ 0 & 0 & 1 \cdots 0 \\ 0 & 0 & 0 \cdots 1 \end{bmatrix} \tag{21}$$

is the identity matrix. An alternate form of (20) is

$$C(z) = A\alpha G_1(z)E(z) + A\alpha\sum_{i=1}^{k} M_i(z)U_i(z), \tag{22}$$

where

$$G_1(z) = \sum_{i=1}^{k} M_i(z), \tag{23}$$

$$M_i(z) = \frac{A_ie^{-a_iT}}{z - e^{-a_iT}}. \tag{24}$$

The pulse transfer function $G_1(z)$ may also be computed from $G(s)$ by use of conventional z-transform tables if the degree of the denominator of $G(s)$ is at least two more than the degree of the numerator of $G(s)$.

For the system shown in Fig. 1 the z transform of the error is

$$E(z) = R(z) - C(z) \tag{25}$$

where $R(z)$ is the z transform of the input to the system.

Substitution of (25) into (22) gives

$$E(z) = E_L(z) + E_N(z), \tag{26}$$

where $E_L(z)$ is the error response of a linear sampled-data system,

$$E_L(z) = \frac{R(z)}{1 + A\alpha G_1(z)}, \tag{27}$$

and $E_N(z)$, an additional signal due to the presence of nonlinearities in the system, is given by

$$E_N(z) = \frac{-A\alpha\sum_{i=1}^{k} M_i(z)U_i(z)}{1 + A\alpha G_1(z)}. \tag{28}$$

It is noted that for small error signals, $E_N(z)$ approaches zero and therefore, under this condition, the error quantity is equal to $E_L(z)$.

Eq. (28) may be expressed in terms of inverse power series as

$$E_N(z) = z^{-1}\sum_{i=1}^{k}\left[\sum_{n=0}^{\infty} w_{in}z^{-n}\right]\left[\sum_{m=0}^{\infty} u_i(mT)z^{-m}\right], \tag{29}$$

where

$$D_i(z) = \frac{A\alpha M_i(z)}{1 + A\alpha G_1(z)} = z^{-1}\sum_{n=0}^{\infty} w_{in}z^{-n} \tag{30}$$

and

$$U_i(z) = \sum_{n=0}^{\infty} u_i(nT)z^{-n}. \tag{31}$$

By definition

$$E_N(z) = \sum_{n=0}^{\infty} e_N(nT)z^{-n}. \tag{32}$$

Comparison of (29) and (32) reveals that

$$\begin{aligned} e_N(0T) &= 0 \\ e_N(T) &= m_{00} \\ e_N(2T) &= m_{01} + m_{10} \\ e_N(3T) &= m_{02} + m_{11} + m_{20} \\ &\vdots \\ e_N(kT) &= \sum_{\alpha=0}^{k-1} m_{\alpha,k-1-\alpha}, \qquad k > 0, \end{aligned} \tag{33}$$

where the m_{ij} are elements of a matrix $\boldsymbol{M}$ defined by

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{D} \tag{34}$$

with the matrices $\boldsymbol{U}$ and $\boldsymbol{D}$ defined by

$$\boldsymbol{U} = \begin{bmatrix} u_1(0T) & u_2(0T) & u_3(0T) & \cdots & u_k(0T) \\ u_1(T) & u_2(T) & u_3(T) & \cdots & u_k(T) \\ u_1(2T) & u_2(2T) & u_3(2T) & \cdots & u_k(2T) \\ u_1(3T) & u_2(3T) & u_3(3T) & \cdots & u_k(3T) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \end{bmatrix}, \tag{35}$$

$$\boldsymbol{D} = \begin{bmatrix} w_{10} & w_{11} & w_{12} & \cdots \\ w_{20} & w_{21} & w_{22} & \cdots \\ w_{30} & w_{31} & w_{32} & \cdots \\ \vdots & \vdots & \vdots & \cdots \\ \vdots & \vdots & \vdots & \cdots \\ w_{k0} & w_{k1} & w_{k2} & \cdots \end{bmatrix}. \tag{36}$$

Since

$$\boldsymbol{M} = \begin{bmatrix} m_{00} & m_{01} & m_{02} & m_{03} & \cdots \\ m_{10} & m_{11} & m_{12} & m_{13} & \cdots \\ m_{20} & m_{21} & m_{22} & m_{23} & \cdots \\ m_{30} & m_{31} & m_{32} & m_{33} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots \end{bmatrix}, \tag{37}$$

the value of $e_N(nT)$ is the sum of the elements in a particular diagonal of the matrix $\boldsymbol{M}$ starting with the upper left-hand corner for $e_N(T)$ and progressing through the array for increasing time. Since $e_L(nT)$ may be determined with (27) through use of linear sampled-data methods, the total error at the sampling instants can then be obtained from the relation

$$e(nT) = e_L(nT) + e_N(nT). \tag{38}$$

Example I

In order to illustrate the application of this method, a particularly simple specific example will be considered.[9] In the control system of Fig. 1, let

$$G(s) = \frac{1}{s(s+1)}$$

and the pulse-width-modulator parameters A, α, and T be chosen as unity. It is desired to determine the response of the system at the sampling instants to step inputs of $\frac{1}{2}$, 1, $1\frac{1}{2}$, 2 and 3.

Solution:

1) Determine A_1 and A_2 $(k=2)$.

$$\frac{1}{s(s+1)} = \frac{A_1}{s} - \frac{A_2}{s+1}; \qquad A_1 = 1, A_2 = -1.$$

[9] It should be noted, however, that no additional complexity whatever is introduced by consideration of an input other than a step function.

2) Determine $G_1(z)$ from (23) and (24).

$$G_1(z) = \frac{0.632z}{(z-1)(z-0.368)}.$$

3) Determine $E_L(z)$ from (27). Since $R(z) = Rz/z - 1$, where R is the magnitude of the step input,

$$E_L(z) = \frac{R(z)}{1+G_1(z)} = \frac{Rz(z-0.368)}{z^2 - 0.736z + 0.368}$$
$$= R[1 + 0.368z^{-1} + 0.097z^{-2} + 0.2065z^{-3}$$
$$- 0.1163z^{-4} - 0.0095z^{-5} + 0.0358z^{-6} + \cdots].$$

4) Determine $D_1(z)$ and $D_2(z)$ from (30).

$$D_1(z) = \frac{z - 0.368}{z^2 - 0.736z + 0.368}$$
$$= z^{-1}[1 + 0.368z^{-1} - 0.097z^{-2} - 0.2065z^{-3}$$
$$- 0.1163z^{-4} - 0.0095z^{-5} + \cdots].$$

$$D_2(z) = \frac{(0.368)(z-1)}{z^2 - 0.736z + 0.368}$$
$$= z^{-1}[0.368 - 0.0097z^{-1} - 0.2065z^{-2}$$
$$- 0.1163z^{-3} - 0.0095z^{-4} + \cdots].$$

5) Form the matrix equation from (34). Note that the matrix $\boldsymbol{D}$ contains numerical values obtained in Step 4 and that the elements of $\boldsymbol{U}$ and $\boldsymbol{M}$ are unknown. Tabulate the values of the linear response obtained in Step 3, providing space for entry of $e_N(nT)$ and $e(nT)$.

n	$e_L(nT)$	$e_N(nT)$	$e(nT)$
0	$e_L(0)$	0	$e_L(0)$
1	$e_L(T)$		
2	$e_L(2T)$		
⋮	⋮		

Now, since $e_N(0)$ is zero from (33),

$$e(0) = e_L(0)$$

as shown in the first entry in the above table. Determine the auxiliary variables $u_i(0)$ from (15a). Then enter these data in row 1 of the matrix $\boldsymbol{U}$, thus allowing determination of m_{00}, which is then entered as $e_N(T)$, since from (33)

$$e_N(T) = m_{00}.$$

Since $e(T)$ is now known, $u_i(T)$ can be determined and entered in the second row of the matrix U, thus allowing computation of the second diagonal m_{10}, m_{01}. Since from (33)

$$e_N(2T) = m_{01} + m_{10},$$

it follows that

$$e(2T) = e_L(2T) + m_{01} + m_{10}.$$

By continuing this process one can determine as many of the $e(nT)$'s as desired.

This process yields the following results for $R=1$:

$$\begin{bmatrix} u_1(0)=0 & u_2(0)=0.718 \\ u_1(T)=0 & u_2(T)=0.245 \\ u_1(2T)=0 & u_2(2T)=0 \\ u_1(3T)=0 & u_2(3T)=-0.08 \\ u_1(4T)=0 & u_2(4T)=-0.04 \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} 1 & 0.368 & -0.097 & -0.2065 & -0.1163 & \cdots \\ 0.368 & -0.097 & -0.2065 & -0.1163 & -0.009 & \cdots \end{bmatrix}$$

$$M = \begin{bmatrix} 0.2645 & -0.0697 & -0.1483 & -0.0836 & -0.0068 \\ 0.0903 & -0.0238 & -0.0506 & -0.0285 & \\ 0 & 0 & 0 & & \\ -0.0294 & -0.0078 & & & \\ -0.0147 & & & & \end{bmatrix}.$$

n	$e_L(nT)$	$e_N(nT)$	$e(nT)$
0	1.000	0	1.000
1	0.368	0.2645	0.632
2	−0.097	0.0205	−0.0765
3	−0.2065	−0.1721	−0.3784
4	−0.1163	−0.1636	−0.2799
5	−0.0095	−0.0422	−0.0517

To determine the closed-loop response for other magnitudes of the step input, repeat the matrix construction of Step 5 with the appropriate values of $e_L(nT)$. For step inputs greater than unity, pulse-width saturation occurs, providing entries in the $u_1(nT)$ column of the U matrix.

The response $c(nT)$ of the system shown in Fig. 1 for step inputs of $R=\frac{1}{2}$, 1, $1\frac{1}{2}$, 2, and 3 is shown in Fig. 2. The output $c(nT)$ is plotted on a normalized scale to facilitate visual examination of the effect of magnitude of a step input on the system response. The line segments joining the values of $c(nT)$ are provided for display purposes only and are not intended to denote the values of $c(t)$ between sampling points.

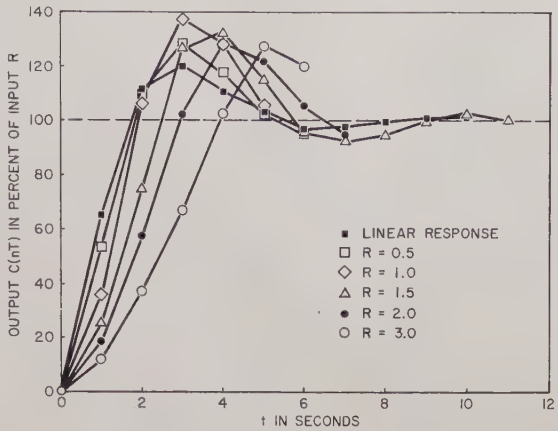


Fig. 2—Response to a step function.

The response curves of Fig. 2 demonstrate the dynamic behavior of a pulse-width-modulated system for varying levels of excitation. For levels of step input less than the pulse-width saturation level ($R=1$), a destabilizing effect with increasing magnitude of the input may be noted. For inputs exceeding the pulse-width saturation level, the decrease in stiffness characteristic of saturation is demonstrated.

Example II

The third order pulse-width-modulated system considered by Nelson[7] is used as a second example. The linear part of this system is

$$G(s) = \frac{0.54(s^2 + 4.85s + 5.58)}{(s+1)(s+2)(s+3)}.$$

Using the higher gain ($\beta=2.5$ and $M=10$ in reference 7), the pulse-width-modulator parameters are

$$\alpha = 1$$
$$A = 4.0.$$

The sampling period T is

$$T = 1.$$

The error response of this system to step inputs of $r(t)=0.5$ and 1.0 as calculated by use of the exact method of analysis presented in this paper are shown in Fig. 3. Again the values of the $c(nT)$ have been joined for display purposes.

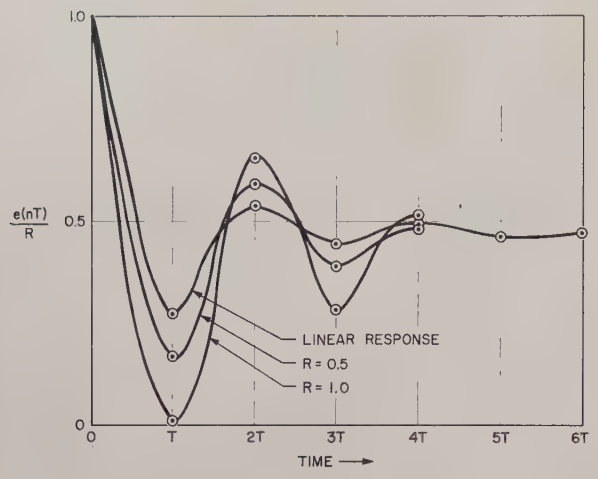


Fig. 3—Error response of a pulse-width-modulated system.

Modified Describing-Function Analysis

A method for studying the stability of pulse-width-modulated systems through use of modified describing functions is presented in this section. In this method the describing-function technique is applied to the equivalent nonlinear sampled-data system which is obtained from the exact difference equations of the pulse-width-modulated system.

The pulse-width-modulated system is defined by the vector equation

$$\mathbf{x}[(n+1)T] = \boldsymbol{G}(T)\mathbf{x}(nT) + A\alpha \operatorname{Sgn} e(nT)\boldsymbol{G}(T-\tau_n)\boldsymbol{h}(\tau_n), \quad (9)$$

where $\boldsymbol{G}(T)$, $\boldsymbol{h}(\tau_n)$ and τ_n are defined in (10), (11), and (14). An auxiliary vector $\boldsymbol{v}$ is defined as

$$\boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \cdot \\ \cdot \\ \cdot \\ v_k \end{bmatrix} \quad (39)$$

where

$$v_i(nT) = e^{-a_iT}\left[\frac{e^{a_i\alpha|e_s(nT)|}-1}{a_i\alpha}\right]\operatorname{Sgn} e(nT). \quad (40)$$

Use of (40) in (9) gives

$$\mathbf{x}[(n+1)T] = \boldsymbol{G}(T)\mathbf{x}(nT) + A\alpha\boldsymbol{v}. \quad (41)$$

The z transform of (41) is

$$z\boldsymbol{X}(z) = \boldsymbol{G}(T)\boldsymbol{X}(z) + A\alpha\boldsymbol{V}(z) \quad (42)$$

provided that $\mathbf{x}(0)=0$. Use of (18) and (19) with (42) gives the z transform of the system output $C(z)$:

$$C(z) = \boldsymbol{A}[z\boldsymbol{I} - \boldsymbol{G}(T)]^{-1}A\alpha\boldsymbol{V}, \quad (43)$$

where $\boldsymbol{A}$ is defined by (19) and $\boldsymbol{I}$ is an identity matrix.

Simplification of (43) gives

$$C(z) = A\alpha\sum_{i=1}^{k}\frac{A_iV_i(z)}{z-e^{-a_iT}}. \quad (44)$$

A block diagram based on (44) and (40) is shown in Fig. 4. This block diagram may be redrawn to include a single nonlinear element in each of the parallel forward paths as shown in Fig. 5. The relocation of the sampler to the output of each nonlinear element is permissible since the nonlinearities are not frequency dependent.

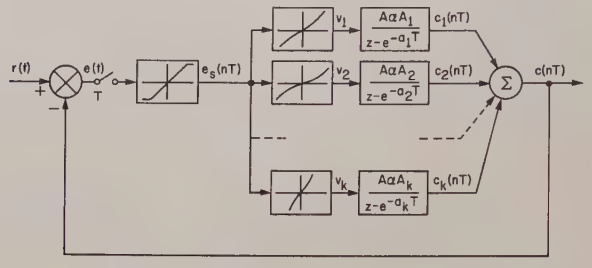


Fig. 4—Block diagram of a pulse-width-modulated control system.

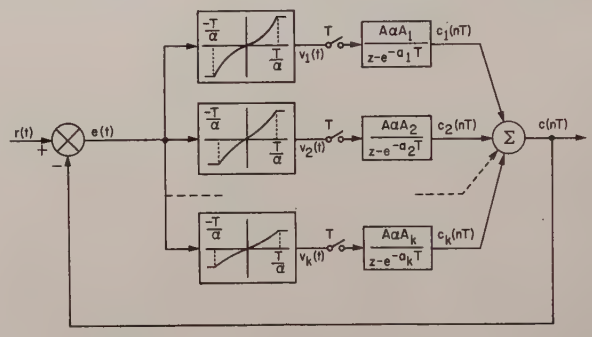


Fig. 5—Block diagram equivalent to that in Fig. 4.

The describing function for each nonlinear element is derived in the Appendix. For $E_m \leq T/\alpha$, the describing function relating an input $e(t)$ and the output $v_i(t)$ is

$$N_i(E_m) = \frac{V_{im}}{E_m} = e^{-a_iT}[1 + 0.425(a_i\alpha E_m) + 0.125(a_i\alpha E_m)^2 + 0.0265(a_i\alpha E_m)^3 + 0.0104(a_i\alpha E_m)^4 + \cdots]; \quad E_m \leq \frac{T}{\alpha}, \quad (45)$$

where

$$e(t) = E_m \operatorname{Sin}\omega t, \quad (46)$$

$$v_i(t) = V_{im}\operatorname{Sin}\omega t + \sum_{k=2}^{\infty} C_k \operatorname{Sin} k\omega t. \quad (47)$$

For magnitudes of E_m greater than the saturation level, the describing function is approximated (see Appendix) by

$$N_i(E_m) = \frac{2}{\pi}\left[\left(\omega t_0 + \frac{\operatorname{Sin} 2\omega t_0}{2}\right) - (1-e^{-a_iT})\cdot\left(\omega t_i + \frac{\operatorname{Sin} 2\omega t_i}{2}\right)\right], \quad E_m \geq T/\alpha, \quad (48)$$

where

$$\omega t_0 = \mathrm{Sin}^{-1}\frac{T/\alpha}{F_m}, \qquad E_m \geq T/\alpha \tag{49}$$

and

$$\omega t_i = \mathrm{Sin}^{-1}\frac{\left[1 - \dfrac{1 - e^{-a_i T}}{a_i T}\right]}{E_m},$$

$$E_m > \left[1 - \frac{1 - e^{-a_i T}}{a_i T}\right]\frac{T}{\alpha}. \tag{50}$$

The output of the control system with a sinusoidal system error of angular frequency ω is obtained by using the describing functions $N_i(E_m)$ with (44). Thus,

$$C(j\omega) = A\alpha\left[\sum_{i=1}^{k}\frac{A_i N_i(E_m)}{e^{j\omega T} - e^{-a_i T}}\right]E(j\omega). \tag{51}$$

For closed-loop operation with $r(t) = 0$,

$$C(j\omega) = -E(j\omega); \tag{52}$$

and from (51) it follows that then

$$1 + G(j\omega, E_m) = 0, \tag{53}$$

where

$$G(j\omega, E_m) = A\alpha\sum_{i=1}^{k}\frac{A_i N_i(E_m)}{e^{j\omega T} - e^{-a_i T}}. \tag{54}$$

Eq. (54) may be expressed in the form

$$G(j\omega, E_m) = \frac{P(j\omega, E_m)}{Q(j\omega)}. \tag{55}$$

The denominator $Q(j\omega)$ is identical to the denominator of the linear pulsed transfer function of the plant $G(s)$ with z replaced with $e^{j\omega T}$. The numerator $P(j\omega, E_m)$ is a polynomial in $e^{j\omega T}$ with coefficients that are functions of the magnitude (E_m) of the sinusoidal error quantity. Stability investigations may be made by plotting a family of curves of $G(j\omega, E_m)$ in the vicinity of the critical point $(-1, 0)$.

Example III

In order to illustrate the application of the describing-function method, a second-order pulse-width-modulated system will be considered. The pulse-width-modulator parameters A, α, and T are all chosen as unity.

$$A = 1$$
$$\alpha = 1$$
$$T = 1.$$

The transfer function of the plant is

$$G(s) = \frac{K}{s(s + 1/2)} = \frac{A_1}{s} - \frac{A_2}{s + 1/2},$$

where

$$A_1 = 2K$$
$$A_2 = -2K.$$

Use of (53) gives

$$1 + 2K\left[\frac{N_1(E_m)}{(e^{j\omega} - 1)} - \frac{N_2(E_m)}{(e^{j\omega} - e^{-1/2})}\right] = 0,$$

$$1 + 2K\,\frac{\{[N_1(E_m) - N_2(E_m)]e^{j\omega} + [N_2(E_m) - e^{-1/2}N_1(E_m)]\}}{(e^{j\omega} - 1)(e^{j\omega} - e^{-1/2})} = 0$$

which is the characteristic equation for the system. This equation may be written as

$$1 + G(j\omega, E_m) = 0,$$

where

$$G(j\omega, E_m) = 2K\,\frac{\{[N_1(E_m) - N_2 E_m)]e^{j\omega} + [N_2(E_m) - 0.6065N_1(E_m)]\}}{(e^{j\omega} - 1)(e^{j\omega} - 0.6065)}.$$

The describing functions $N_1(E_m)$ and $N_2(E_m)$ are from (45) and (48):

$$N_1(E_m) = 1, \qquad E_m \leq 1$$

$$N_1(E_m) = \frac{2}{\pi}\left(\omega t_0 + \frac{\mathrm{Sin}\ 2\omega t_0}{2}\right), \qquad E_m > 1,$$

where

$$\omega t_0 = \mathrm{Sin}^{-1}\left(\frac{1}{E_m}\right)$$

$$N_2(E_m) = 0.6065[1 + 0.2125E_m + 0.031{E_m}^2 + 0.0033{E_m}^3 + 0.0006{E_m}^4 + \cdots], \qquad E_m \leq 1,$$

$$N_2(E_m) = \frac{2}{\pi}\left[\left(\omega t_0 + \frac{\mathrm{Sin}\ 2\omega t_0}{2}\right) - (1 - e^{-1/2})\cdot\left(\omega t_2 + \frac{\mathrm{Sin}\ 2\omega t_2}{2}\right)\right], \qquad E_m > 1,$$

where

$$\omega t_0 = \mathrm{Sin}^{-1}\left(\frac{1}{E_m}\right), \qquad E_m > 1$$

and

$$\omega t_2 = \mathrm{Sin}^{-1}\left(\frac{0.213}{E_m}\right), \qquad E_m > 0.213.$$

Graphs of $N_1(E_m)$ and $N_2(E_m)$ are presented in Fig. 6.

Nyquist diagrams of $G(j\omega, E_m)/2K$ for constant values of $E_m = 0$, 0.2, 0.4, 0.6, 0.8, 1.0 are plotted in Fig. 7. For $K = 1.72$ the critical point $(-1/2K)$ is -0.29, and the system is stable for sinusoidal error magnitudes less than approximately 0.7. For larger error magnitudes, the system is unstable. An unstable limit cycle exists for a sinusoidal error magnitude in the vicinity of 0.7 at a frequency of approximately 0.2 cps.

Nyquist diagrams of $G(j\omega, E_m)/2K$ for constant values of $E_m = 1.0$, 1.4, 1.6, 1.8, 2.0 are presented in Fig. 8. The intersection of the graph of $G(j\omega, E_m)/2K$ with the critical point $(-1/2K, 0)$ for $K = 1.72$ establishes the existence of a stable limit cycle with a magnitude of $E_m = 1.8$ and a frequency of 0.16 cps.

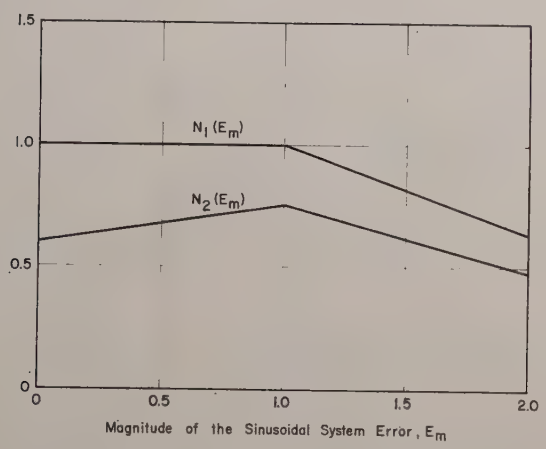


Fig. 6—Describing functions used in Example 3.

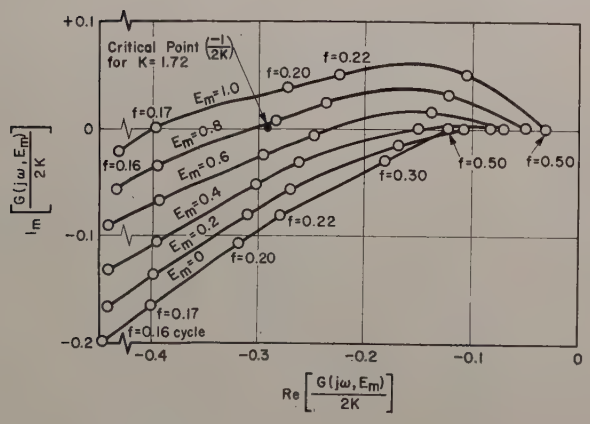


Fig. 7—Nyquist diagrams for a pulse-width-modulated control system (below saturation).

A confidence check on the describing-function results may be made by determining the exact response of the system to step inputs of suitable magnitudes. The error response of the system being studied in this example for $K = 1.72$ and step inputs of 0.6, 0.8, and 2.0 was determined by use of the exact method presented in this paper and is plotted in Fig. 9. It is noted that for a step input of 0.6 the system is stable. For a step input of 0.8 the system is unstable, with a natural frequency of approximately 0.2 cps. This compares well with the predicted unstable limit cycle for $E_m = 0.7$ at a frequency of 0.2 cps.

For a step input of 2.0, the error response fits well to a sinusoidal curve with a magnitude of 1.8 at 0.16 cps,

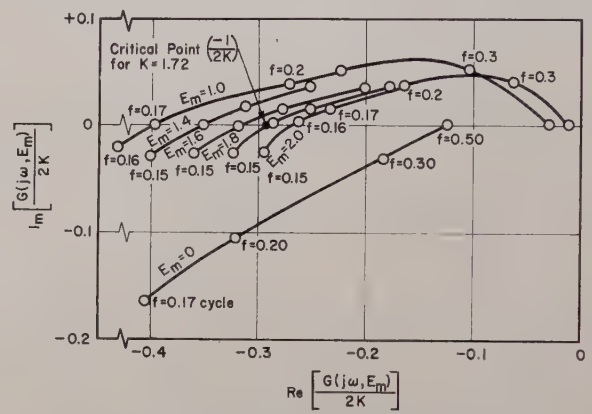


Fig. 8—Nyquist diagrams for a pulse-width-modulated control system (above saturation).

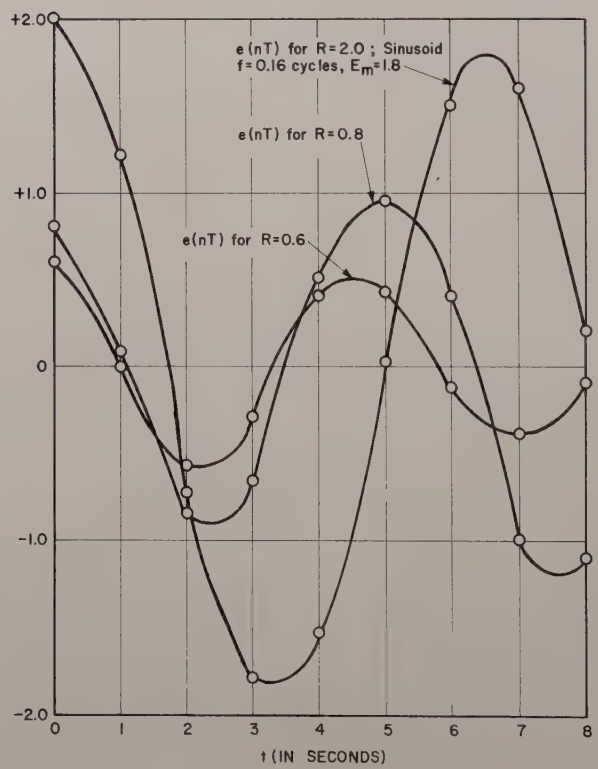


Fig. 9—Transient response of a pulse-width-modulated control system.

which compares well with the predicted stable limit cycle.

It should be observed that the data displayed in Fig. 9 are valid only at the sampling instants and that the curves joining the $c(nT)$ are for display purposes only.

Conclusions

The general objective of this paper has been to present a method for studying the performance of closed-loop pulse-width-modulated control systems. An exact analytical method was presented for the determination of the response of such systems to arbitrary inputs. The application of this method to representative second- and third-order systems demonstrates the effect of pulse-width on the dynamic performance of the system.

The combined use of the modified describing-function method and the exact-response method provides a basis for system design. The describing-function approach may be used to reveal the general stability boundaries and limit-cycle conditions, thereby providing system design criteria. The validity of the describing-function design may then be substantiated by system analysis in the time domain by the use of the exact system response method.

The use of describing functions is always limited by the assumption of adequate filtering of the harmonics. It might be suspected therefore that the results obtained on the basis of Fig. 5 are not very accurate. However, an exact analysis has revealed that, in the examples considered, adequate accuracy is obtained. This can perhaps more readily be understood after recognizing that the diagram in Fig. 5 is equivalent to one in which the original pulse-width modulator is simply replaced by its conventional describing function.

Appendix

The describing function for the nonlinear element that converts $e(t)$ into $v_i(t)$ is derived in this Appendix. The defining equation for $v_i(t)$ is

$$v_i(t) = e^{-a_iT}\left[\frac{e^{a_i\alpha|e_s(t)|}-1}{\alpha a_i}\right], \tag{56}$$

where

$$e_s(t) = \begin{cases} e(t), & |e(t)| \le \dfrac{T}{\alpha} \\ \dfrac{T}{\alpha}\operatorname{Sgn} e(t), & |e(t)| > \dfrac{T}{\alpha}. \end{cases} \tag{57}$$

For

$$e(t) = E_m \operatorname{Sin} \omega t, \tag{58}$$

the output v_i is

$$v_i(t) = e^{-a_iT}\left[\frac{e^{a_i\alpha E_m \operatorname{Sin}\omega t}-1}{\alpha a_i}\right], \qquad 0 \le \omega t < \pi \quad \text{and} \quad E_m < \frac{T}{\alpha}. \tag{59}$$

Expansion of the exponential term in an infinite power series gives

$$v_i(t) = \frac{e^{-a_iT}}{\alpha a_i}\left[X \operatorname{Sin}\omega t + \frac{X^2}{2!}\operatorname{Sin}^2\omega t + \frac{X^3}{3!}\operatorname{Sin}^3\omega t + \cdots\right], \tag{60}$$

where $X = a_i\alpha E_m$.

For inputs $e(t)$ below the saturation level, the describing function $N_i(E_m)$ is

$$N_i(E_m) = \frac{4}{\pi}\frac{e^{-a_iT}}{\alpha a_i}\int_0^{\pi/2}\left[\sum_{n=1}^{\infty}\frac{X^n}{n!}\operatorname{Sin}^{n+1}\omega t\right]d\omega t. \tag{61}$$

Integration of (61) yields

$$N_i(E_m) = e^{-a_iT}E_m[1 + 0.425(a_i\alpha E_m) + 0.125(a_i\alpha E_m)^2 + 0.0265(a_i\alpha E_m)^3 + 0.0104(a_i\alpha E_m)^4 + \cdots] \tag{62}$$

for $E_m \le T/\alpha$.

For inputs $e(t)$ exceeding the saturation level, an approximation to the describing function $N_i(E_m)$ may be conveniently obtained by approximating the v_i vs e in the interval $0 < e < T/\alpha$ with two straight line segments. The line segments are chosen to fit the $v_i - e$ characteristic at $e = 0$ and $e = T/\alpha$. With this approximation, v_i is

$$v_i(t) = e_s(t) - (1 - e^{-a_iT})e_{s1}(t), \tag{63}$$

where

$$e_s(t) = \begin{cases} e(t), & |e(t)| < \dfrac{T}{\alpha} \\ \dfrac{T}{\alpha}\operatorname{Sgn} e(t), & |e(t)| > \dfrac{T}{\alpha} \end{cases} \tag{64}$$

and

$$e_{s1}(t) = \begin{cases} e(t), & |e(t)| \le \left[\dfrac{1}{1-e^{-a_iT}} - \dfrac{1}{a_iT}\right]\dfrac{T}{\alpha} \\ \dfrac{T}{\alpha}\left[1 - \dfrac{1-e^{-a_iT}}{a_iT}\right]\operatorname{Sgn} e(t), & |e(t)| > \left[\dfrac{1}{1-e^{-a_iT}} - \dfrac{1}{a_iT}\right]\dfrac{T}{\alpha}. \end{cases} \tag{65}$$

Since (63), (64), and (65) define saturating linear functions, the describing function $N_i(E_m)$ is given by

$$N_i(E_m) \cong \frac{2}{\pi}\left[\left(\omega t_0 + \frac{\operatorname{Sin} 2\omega t_0}{2}\right) - (1 - e^{-a_i T})\left(\omega t_i + \frac{\operatorname{Sin} 2\omega t_i}{2}\right)\right], \quad (66)$$

where

$$\omega t_0 = \begin{cases} \operatorname{Sin}^{-1}\left(\dfrac{T/\alpha}{E_m}\right), & E_m \geq \dfrac{T}{\alpha} \\ \dfrac{\pi}{2}, & E_m < \dfrac{T}{\alpha} \end{cases} \quad (67)$$

and

$$\omega t_i = \begin{cases} \operatorname{Sin}^{-1} \dfrac{\left[\dfrac{1}{1 - e^{-a_i T}} - \dfrac{1}{a_i T}\right]\dfrac{T}{\alpha}}{E_m}, & E_m > \left[\dfrac{1}{1 - e^{-a_i T}} - \dfrac{1}{a_i T}\right]\dfrac{T}{\alpha} \\ \dfrac{\pi}{2}, & E_m < \left[\dfrac{1}{1 - e^{-a_i T}} - \dfrac{1}{a_i T}\right]\dfrac{T}{\alpha}. \end{cases} \quad (68)$$

The exact describing function for $E_m > T/\alpha$ may be obtained by using (60) for $e(t) \leq T/\alpha$ and (56) for $e(t) > T/\alpha$.

The integral form of the describing function is

$$N_i(E_m) = \frac{4}{\pi}\frac{e^{-a_i T}}{\alpha a_i}\left[\int_0^{\omega t_a}\left(\sum_{n=1}^{\infty}\frac{X^n}{n!}\operatorname{Sin}^{n+1}\omega t\right)d\omega t + (e^{a_i T} - 1)\left(\frac{\pi}{2} - \omega t_a\right)\right], \quad (69)$$

where

$$X = a_i \alpha E_m \quad (70)$$

and

$$\omega t_a = \begin{cases} \operatorname{Sin}^{-1}\dfrac{T/\alpha}{E_m}, & E_m > \dfrac{T}{\alpha} \\ \dfrac{\pi}{2}, & E_m < \dfrac{T}{\alpha}. \end{cases} \quad (71)$$

The approximation to the describing function, given in (66), was used in the illustrative example to avoid the tedious computation of the integral in (69).

Effects of Quantization on Feedback Systems with Stochastic Inputs*

R. KRAMER†, MEMBER, IRE

Summary—**An approximate analysis of the effects of quantization in a feedback system is made. The system input is a Gaussian random signal. The error autocorrelation as a function of the quantizer box size is the goal of the analysis. The approximation lies in the assumption that certain error joint distributions are Gaussian. In the limit as the quantizer box size approaches zero, these distributions do become Gaussian.**

On the basis of the approximation, a nonlinear integral equation relating the error autocorrelation to the system parameters is developed. An iteration procedure for successive approximations to the solution is outlined, and several examples are presented. Finally, experimental results obtained on a digital computer are shown.

* Received by the PGAC, August 4, 1960; revised manuscript received, February 6, 1961. This paper is based upon a thesis submitted in partial fulfillment of the Sc.D. degree in the Dept. of Elec. Engrg., at Mass. Inst. Tech. in June, 1959. The work was supported jointly by contracts No. AF-33(316)-5477 under the sponsorship of the USAF and No. DA-19-020-ORD-4637 under the sponsorship of the U. S. Army.

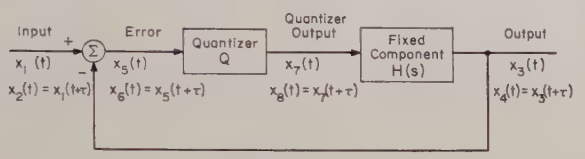
† Electronics Systems Lab., Mass. Inst. Tech., Cambridge, Mass.

I. Introduction

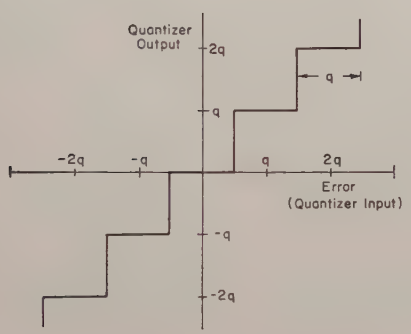
IN THE FIELD of feedback control systems, linear systems are quite well understood; nonlinear systems are not yet so favored. There is no general theory of nonlinear systems analysis; rather, each particular type of nonlinearity seems to demand a special analysis technique. The desire to analyze nonlinear sys-

tems has two motives beyond the academic one of desiring to extend our understanding of systems in general. These motives lie first in the fact that many systems have inherent, dominant nonlinearities, and second in the possibility that the deliberate introduction of a nonlinearity may result in performance better than could be attained with a strictly linear system. The investigation reported here is of a system of the former class—the nonlinearity is essentially inherent in the system.

The nonlinearity in the system consists of a quantizer in the error channel. A quantizer is a device with no memory, whose output can exist only at certain discrete levels determined by the input. The system and the quantizer gain function with which we are concerned are shown in Fig. 1. The problem is that of determining



(a)



(b)

Fig. 1—Configuration of feedback system with quantizer. (a) System block diagram. (b) Quantizer gain function.

the effects of a quantizer in the error channel of a feedback system. There are two general classes of effects: effects on the over-all, absolute stability of the system and effects on the detailed performance of the system. It is the latter that is of concern here. We assume that the absolute stability of the system has been assured—including the effects of the nonlinearity—and that a more detailed examination of the performance is desired. To this end, a Gaussian random signal has been chosen to test the effects of the quantizer on the performance of the system.

There are many systems in which a quantizer in the error channel is either inherent or desirable. The most obvious example of a system in which quantization is present is a digital computer. The finite number of digits which are handled represents a quantization of the data. Of course, most computers handle a large number of digits, so that the effects of quantization are generally quite small. However, in cases where the computer is interacting with the real (analog) world and, in particular, where this interaction is included within a feedback loop, there is pertinency in the question of how many digits must be carried in order to ensure satisfactory over-all performance. The effects of the quantization in time *i.e.*, sampling, also inherent in digital systems are not included here. Sampled data systems have already received wide attention [1], [2]. Quantization also is pertinent in systems in which the error sensor has inherently a quantized output. Such a sensor would be a digital camera in which a finite number of photo cells are used to measure the angular offset of a beam of light [3].

II. Analysis

The analysis of the system shown in Fig. 1 begins with some general relations:

$$\phi_{56}(\tau) = \phi_{12}(\tau) + \phi_{34}(\tau) - \phi_{14}(\tau) - \phi_{14}(-\tau)^1 \quad (1)$$

$$\phi_{16}(\tau) = \phi_{12}(\tau) - \phi_{14}(\tau) \quad (2)$$

$$\phi_{14}(\tau) = \int_{-\infty}^{\infty} d\sigma h(\sigma)\phi_{18}(\tau - \sigma) \quad (3)$$

$$\phi_{34}(\tau) = \int_{-\infty}^{\infty} d\sigma \int_{-\infty}^{\infty} dv h(v)h(v + \sigma)\phi_{78}(\tau - \sigma). \quad (4)$$

Of these relations, the first two are general correlation relations among the variables at the system summing (error-determining) point. The second two are relations among correlations of variables associated with a linear, dynamic element whose impulse response is $h(\tau)$ with a corresponding transform $H(s)$. These equations are not sufficient to allow the solution for the error autocorrelation $\phi_{56}(\tau)$; indeed, these relations do not involve the quantizer at all. In order to solve these equations, two more relations are required:

$$\phi_{78}(\tau) = g[\phi_{56}(\tau)] \quad (5)$$

$$\phi_{18}(\tau) = f[\phi_{16}(\tau)]. \quad (6)$$

These relationships depend directly upon certain joint distributions. Specifically, the autocorrelation relationship of (5) depends on the joint distribution of the error $P_{56}(X_5, X_6; \tau)$. The cross-correlation relation of (6) depends upon the joint distribution between the error and the system input $P_{16}(X_1, X_6; \tau)$. These distributions have to be known before the required relationships can be determined. But these distributions are the very solutions we are trying to obtain; and at the present state of the art, it does not appear possible to solve for them. However, we know that, as the quantizer box-size q approaches zero, the system approaches

[1] The subscripts refer to the signals involved in the correlation or distribution. The subscript 1 refers to the input $x_1(t)$ and the subscript 2 refers to the input variable $x_2(t)$ which is the signal x_1 shifted in time by τ seconds. Similarly, $x_3(t)$ is the system output and $x_4(t)$ is this same output shifted in time. The other variables are treated similarly. See Fig. 1.

linearity and these distributions become Gaussian for a Gaussian input. Consequently, as a first approximation, it is assumed that these distributions are Gaussian in the expectation that this will permit an extension of the analysis somewhat beyond the strictly linear case.

On the basis of this approximation, the relations of (5) and (6) and their dependence on quantizer box size q can be determined analytically. The derivation of the first relation is fundamentally that of Widrow [4]. The basic analysis is perfectly general and applicable to any specified joint distribution at the quantizer input. In order to determine the actual autocorrelation at the quantizer output in terms of that at the input, however, the specific distribution must be known or assumed. For the assumed Gaussian distribution, we obtain, as shown in Appendix I,

$$\phi_{78}(\tau) = \phi_{56}(\tau)\left[1 + 4\sum_{n=1}^{\infty}(-1)^n \exp - 1/2\left(\frac{2\pi\sigma_5}{q}\right)^2 n^2\right] + 2\sigma_5^2\left(\frac{q}{2\pi\sigma_5}\right)^2 \sum_{m,n=1}^{\infty}\frac{(-1)^{m+n}}{mn} \cdot \left\{\exp - 1/2\left(\frac{2\pi\sigma_5}{q}\right)^2 [m^2 + n^2 - mn\rho_{56}(\tau)] - \exp - 1/2\left(\frac{2\pi\sigma_5}{q}\right)^2 [m^2 + n^2 + mn\rho_{56}(\tau)]\right\} \quad (7)$$

where

$$\phi_{56}(\tau) = \sigma_5^2\rho_{56}(\tau)$$

with σ_5^2 being the variance of the random variable x_5.

The dependency of this relation upon the quantizer box size appears in the ratio (q/σ_5) which is the relative box size; Fig. 2 shows a plot of this function for several values of (q/σ_5). This function of (7) may be expanded into a proportional term and a nonlinear term:

$$\phi_{78}(\tau) = a\phi_{56}(\tau) + \sigma_5^2 f[\rho_{56}(\tau)] \quad (8)$$

where the coefficient a of the linear term is

$$a = \frac{d}{d\rho_{56}}\left(\frac{\phi_{78}}{\sigma_5^2}\right)\Bigg|_{\rho_{56}=0} = 1 + 4\sum_{n=1}^{\infty}(-1)^n \exp - 1/2\left(\frac{2\pi\sigma_5}{q}\right)^2 n^2 + 4\sum_{m,n=1}^{\infty}(-1)^{m+n} \exp - 1/2\left(\frac{2\pi\sigma_5}{q}\right)^2 (m^2 + n^2) \quad (9)$$

as shown in Appendix I. This particular expansion of the autocorrelation functional relationship of (7) is particularly convenient for the solution of the system equations.

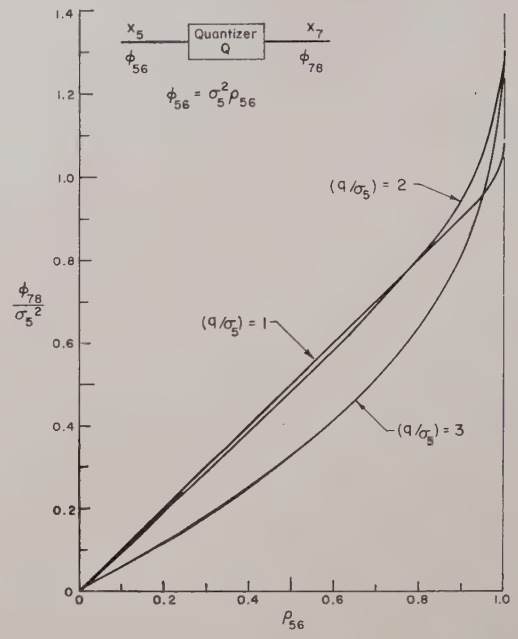


Fig. 2—Functional relation between quantizer input and autocorrelations.

The cross-correlation relationship (6) has been studied by a number of people [5]–[7] for nonlinear gains in general. The specific form that it takes depends upon the properties of the joint distribution $P_{16}(X_1, X_6; \tau)$. For distributions having the property called "separability" by Nuttall [5], this cross-correlation relationship reduces to one of proportionality:

$$\phi_{18}(\tau) = K_Q\phi_{16}(\tau). \quad (10)$$

In order to evaluate the constant of proportionality, the specific distribution $P_{16}(X_1, X_6; \tau)$ must be known. With the assumption of a Gaussian distribution, the value of K_Q for a quantizer type of nonlinear gain becomes

$$K_Q = \frac{1}{\sigma_5^3\sqrt{2\pi}}\sum_{n=-\infty}^{\infty} nq\int_{(n-1/2)q}^{(n+1/2)q} x_5 \cdot \exp\left[-1/2\left(\frac{x_5}{\sigma_5}\right)^2\right] dx_5 \quad (11)$$

as shown in Appendix II.

This constant K_Q is also a function of (q/σ_5); the relationship is plotted in Fig. 3. Furthermore, it is shown in Appendix III that the constant a of (9) and the K_Q of (11) are related:

$$a = K_Q^2. \quad (12)$$

Eqs. (1)–(4) with (8) and (10) may be reduced to the following nonlinear integral equation relating the error autocorrelation $\phi_{56}(\tau)$ to the input autocorrelation $\phi_{12}(\tau)$, the fixed component dynamics $h(t)$, and the

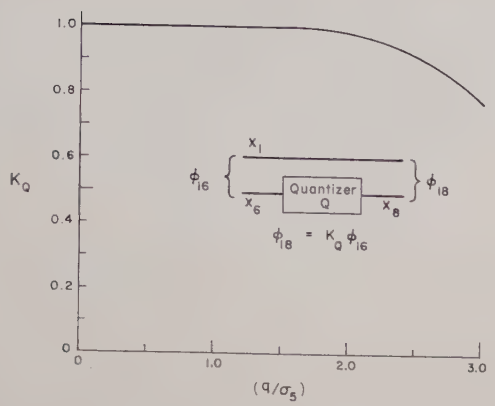


Fig. 3—Quantizer cross-correlation proportionality constant.

quantizer relative box size (q/σ_5):

$$\phi_{56}(\tau) = \phi_{12}(\tau) - \int_{-\infty}^{\infty} dv g(v)[\phi_{12}(\tau - v) + \phi_{12}(\tau + v)] + K_Q^2 \int_{-\infty}^{\infty} d\sigma \int_{-\infty}^{\infty} dv h(v)h(v + \tau - \sigma)\phi_{56}(\sigma) + \sigma_5^2 \int_{-\infty}^{\infty} d\sigma \int_{-\infty}^{\infty} dv h(v)h(v + \tau - \sigma)f[\rho_{56}(\sigma)] \quad (13)$$

where

$$g(v) = \mathfrak{F}^{-1}\left[\frac{K_Q H(s)}{1 + K_Q H(s)}\right]. \quad (14)$$

Eq. (13) then is the desired relationship between the system error autocorrelation $\phi_{56}(\tau)$ and the input autocorrelation $\phi_{12}(\tau)$, the system dynamics $h(v)$, and the quantizer through K_Q and $f[\rho_{56}]$.

The input-output correlation may also be derived from these simultaneous equations and is

$$\phi_{14}(\tau) = \int_{-\infty}^{\infty} dv g(v)\phi_{12}(\tau - v). \quad (15)$$

The technique for solving the integral equation (13) for $\phi_{56}(\tau)$ is one of successive approximation. In its most obvious form, the method consists of taking an approximate solution, inserting it into the right-hand side of (13), and then using the resulting left-hand side as the next approximation. This process may be stated in a somewhat more formal and more general manner as follows.

The nth approximation $\phi_n(\tau)$ consists of a sum of terms

$$\phi_n(\tau) = \phi_0(\tau) + \sum_{k=1}^{n} \Delta_k(\tau) \quad (16)$$

where $\Delta_k(\tau)$ is a general correction term which we must yet define. The amount (a function of time τ) by which $\phi_n(\tau)$ fails to satisfy the integral equation is a measure of the quality of the nth approximation to the solution. We call this amount, or measure, the difference function $\delta_n(\tau)$ defined as

$$\delta_n(\tau) = \phi_{12}(\tau) - \int_{-\infty}^{\infty} dv g(v)[\phi_{12}(\tau - v) + \phi_{12}(\tau + v)] + K_Q^2 \int_{-\infty}^{\infty} d\sigma \int_{-\infty}^{\infty} dv h(v)h(v + \tau - \sigma)\phi_n(\sigma) + \sigma_{5n}^2 \int_{-\infty}^{\infty} d\sigma \int_{-\infty}^{\infty} dv h(v)h(v + \tau - \sigma)f[\rho(\sigma)] - \phi_n(\tau). \quad (17)$$

From this function, we may define a general correction term

$$\Delta_{n+1}(\tau) = -C(\tau)\delta_n(\tau). \quad (18)$$

In this correction term, if $C(\tau) = -1$, then the resulting sequence of approximations $\phi_n(\tau)$ will be identical to that obtained as described above where the $(n+1)$st approximation is obtained by substituting the nth approximation into the right-hand side of the defining integral equation (13).

If we select as a first approximation $\phi_0(\tau)$, the solution to the linear integral equation obtained by ignoring the last, nonlinear term of (13), we achieve certain simplifications. For one thing, as the quantizer box size gets smaller, this nonlinear term approaches zero, and the first approximation $\phi_0(\tau)$ becomes the exact solution. Secondly, the expression for $\delta_n(\tau)$ becomes

$$\delta_n(\tau) = K_Q^2 \int_{-\infty}^{\infty} d\sigma \int_{-\infty}^{\infty} dv h(v)h(v + \tau - \sigma) \sum_{k=1}^{n} \Delta_k(\sigma) + \sigma_{5n}^2 \int_{-\infty}^{\infty} d\sigma \int_{-\infty}^{\infty} dv h(v)h(v + \tau - \sigma)f[\rho_n(\sigma)] - \sum_{k=1}^{n} \Delta_k(\tau). \quad (19)$$

Wagner [8], in an article on iterative solutions to a Fredholm equation of the second kind, suggests that this factor $C(\tau)$ be

$$C(\tau) = \frac{-1}{1 - K_Q^2 \int_{-\infty}^{\infty} d\sigma \int_{-\infty}^{\infty} dv h(v)h(v + \tau - \sigma)}. \quad (20)$$

Notice that the double integral amounts simply to the square of the area under $h(\tau)$ and is independent of τ. Furthermore, the area under $h(\tau)$ is simply the dc gain of the fixed element in the system. Consequently,

$$C(\tau) = \frac{-1}{1 - (K_Q K_f)^2} \quad (21)$$

where K_f is the aforementioned dc gain of the fixed element. Wagner suggests that, when $(K_Q K_f)^2$ is in the neighborhood of one, that $C(\tau) = -1$ be used instead. [At $K_Q^2 K_f^2 = 1$, of course, the value of $C(\tau)$ from (21) goes to infinity.]

This general procedure as summarized by (16), (18), and (19) is used to solve the integral equation for the error autocorrelation. It must be pointed out here that this technique breaks down when the fixed component $h(\tau)$ of the system contains an integration. The difficulty that arises is that, if the output variance σ_3^2 is to be finite and if the fixed component $h(t)$ contains an integration, the input to $h(t)$ must have an autocorrelation with zero area. If the system without quantization has a finite output variance, there is no reason to expect this variance to go to infinity as quantization is introduced. Therefore, the autocorrelation $\phi_{78}(\tau)$ of the input to the fixed element must have zero area in order to result in a finite output variance in a system containing an integration. This autocorrelation is related to the error autocorrelation $\phi_{56}(\tau)$ by

$$\phi_{78}(\tau) = a\phi_{56}(\tau) + \sigma_5^2 f[\rho_{56}(\tau)]. \tag{8}$$

In order to ensure that, at each stage of approximation to $\phi_{56}(\tau)$, the autocorrelation $\phi_{78}(\tau)$ has zero area, the approximate correlations $\phi_n(\tau)$ must be so constrained. Thus far, we have found no way to introduce this constraint in the iterative procedure.

III. Examples

The analysis technique described above was applied to several examples to gain some insight into the effects of quantization in some typical situations, as well as to gain some experience in the solution procedure. The examples which were studied were:

Case 1:

$$\phi_{12}(\tau) = \sigma_1^2 e^{-\gamma|\tau|}; \qquad H(s) = \frac{\gamma/2}{s + \gamma/2}; \qquad (q/\sigma_5) = 1, 3.$$

Case 2:

$$\phi_{12}(\tau) = \sigma_1^2(1 + \gamma|\tau|)e^{-\gamma|\tau|};$$

$$H(s) = \frac{\gamma}{s + \gamma}; \qquad (q/\sigma_5) = 1, 2, 3.$$

Case 3:

$$\phi_{12}(\tau) = \sigma_1^2(1 + \gamma|\tau|)e^{-\gamma|\tau|};$$

$$H(s) = \frac{10\gamma}{s + \gamma}; \qquad (q/\sigma_5) = 1, 3.$$

These cases were selected to point up various aspects of the solution procedure.

A number of interesting points were revealed by the analytical study of these examples. The effects of the quantizer on the error autocorrelation depend to a large extent upon the behavior of the error autocorrelation in the neighborhood of $\tau = 0$. As may be seen from Fig. 2, the relationship between the quantizer output correlation and that of the input shows a definite peaking near $\rho_{56} = 1$, which corresponds to $\tau = 0$. When the error autocorrelation drops sharply from its initial value as τ increases, the effect of the nonlinear term in the integral equation (13) is very small. It is when the initial slope of the error autocorrelation is low and the normalized correlation $\rho_{56}(\tau)$ remains near unity for larger values of τ, that this nonlinear term becomes important. The behavior of $\rho_{56}(\tau)$ around the origin is primarily dependent upon the behavior of the input autocorrelation around the origin and also upon the relationship between the system closed-loop bandwidth and the bandwidth of the input power-density spectrum. When the input correlation has low slope at the origin, so too will the error autocorrelation. However, when the system bandwidth is large compared to that of the input, the slope of the error autocorrelation, as τ increases, decreases rapidly from its value at $\tau = 0$. Thus, when the error autocorrelation decreases rapidly from its value at the origin by virtue of the input autocorrelation and/or the system bandwidth, the effects of quantization in general are reduced.

There are two sources of influence of the quantizer on the error autocorrelation. One is through the effective gain reduction introduced by K_Q, and the other is through the nonlinear term in the integral equation. The examples seemed to indicate that, for $(q/\sigma_5) \leq 2$, the contribution of the nonlinear term is minor. Calculations show that the difference function $\delta_0(\tau)$ associated with the initial approximation $\phi_0(\tau)$, which is directly dependent upon K_Q, is reduced in the subsequent iterations principally through slight changes in the error variance. The normalized autocorrelation undergoes a negligible change with these iterations. For $(q/\sigma_5) = 3$, the iteration procedure results in larger changes in the error variance and also in the normalized correlation, though the changes in the latter still seem to be small.

The convergence of the iteration procedure seems satisfactory. Only a few iterations are necessary to observe the trend of the approximation and reduce the difference function significantly. For cases where the loop gain is greater than one, the convergence factor is necessary for absolute stability of the iteration procedure. This convergence factor has a strong influence on the convergence characteristics of the iteration procedure.

In Fig. 7 are shown the plots from a typical iterative calculation. The convergence characteristics of the iterative solution are seen in the plots of the successive difference functions $\delta_n(\tau)$. Note how small a change occurred in the error variance σ_{5n}^2/σ_1^2 in this sequence;

actually, there were changes in ρ_n which were of the same order of magnitude. The other examples showed similar trends, although the changes in the error variance were generally larger; the normalized error autocorrelation tended, in general, to change very little.

IV. Experimental Program

An experimental program was carried out as a means of evaluating the analysis. As indicated above, the analysis was based on the approximation that certain joint distributions in the system were Gaussian. As the quantizer box size becomes zero, this approximation becomes exact. The objective of the experimental program was to give some indication of the range of quantizer box sizes for which the correlation relations of (7)–(10) (which are based on this assumption) are still approximately correct.

The experimental work consisted of generating a random signal of appropriate autocorrelation, exciting a feedback system with it, and measuring the appropriate correlation functions. The experimentation was carried out digitally on an IBM 704 computer. The specific objectives of the experiment were:

1) Direct measurement of the error autocorrelation for comparison with the calculated functions.
2) Measurement of the quantizer output autocorrelation in conjunction with the quantizer input autocorrelation, so that the functional relationship
$$\phi_{78}(\tau) = g[\phi_{56}(\tau)] \tag{5}$$
could be determined.
3) Measurement of the cross-correlations $\phi_{16}(\tau)$ and $\phi_{18}(\tau)$ so that the relationship
$$\phi_{18}(\tau) = f[\phi_{16}(\tau)] \tag{6}$$
could be determined.

The accuracy of results desired was about 5 per cent or better. An approximate analysis of the errors introduced by sampling and by using a finite length of data for the correlations was made. When the sampling period is short enough, the standard deviation σ_M of of the error in the measured variance reduces to

$$\sigma_M = \sigma_x^2 \sqrt{\frac{2}{\alpha T}} \tag{22}$$

where

σ_x^2 = the true value of the variance
T = the length of sample correlated
$1/\alpha$ = the time constant of the simple exponential correlation function assumed in the analysis.

This is based on the error analysis shown in Appendix IV. On this basis, αT must be about 2000 (a length of data 2000 times the characteristic time of the true correlation) in order that the standard deviation of the measured variance be about 3 per cent of the true value of the variance. This large quantity of data dictated the use of a digital computer for the experimentation.

The digital simulation of the system required two filters: one to filter the output of a standard, uncorrelated, Gaussian random-number generator to provide the appropriate input signal autocorrelation; and a second to simulate the fixed component $H(s)$. The simulation approach was first to represent the dynamics in terms of a feedback system using only integrators for dynamic elements. Then the integrators were replaced by an approximate digital integration using a simple rectangular integration rule. The resulting equation was used as the digital simulation of the dynamic components in the experiment. The computation interval was arbitrarily selected to be $1/100\alpha$ in the expectation that this would be sufficiently short so that, with the rectangular integration rule, satisfactory simulation would be achieved. This interval choice was then checked by computing the input autocorrelation function. In Fig. 4 is shown a comparison of the desired input autocorrelation and the measured ($\alpha T = 2000$) autocorrelation; the agreement is excellent, which implies, among other things, that the filter simulation is satisfactory.

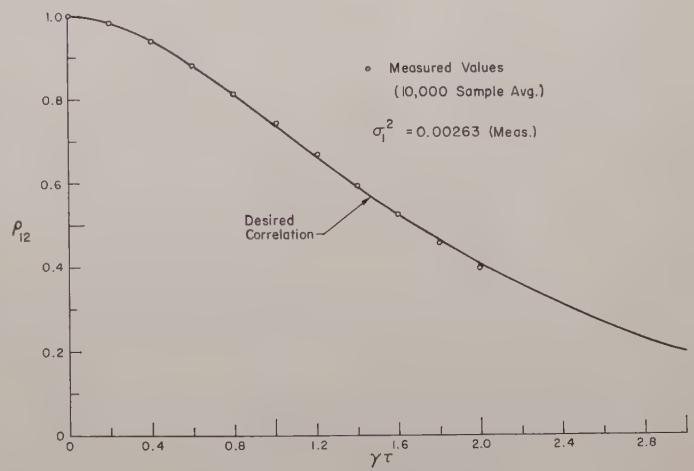


Fig. 4—Comparison of desired and experimental input autocorrelation.

The experimental results show that, for $(q/\sigma_5) \leq 2$, the approximations are quite good; beyond this value, the approximations rapidly become very poor. In Fig. 5 is shown a comparison of the approximate analytical with the experimental relationship between the quantizer output autocorrelation and that of the quantizer input. Notice that, for $(q/\sigma_5) = 1$, the agreement is excellent; and that, for $(q/\sigma_5) = 2$, while there is a quite definite deviation, the analytic function is probably still a very good approximation. Although for $(q/\sigma_5) = 3$ the agree-

ment is poor, yet the experimental relationship shows general characteristics similar to the analytic relationship implying that a better approximation to this functional relationship may be possible. A comparison of the approximate analytic relation with the experimental relation between the quantizer cross-correlations is shown in Fig. 6. Again, the general pattern of results observed for the autocorrelations is repeated. The analytic approximations are in reasonably good agreement with the experimental results for $(q/\sigma_5) \leq 2$; and beyond this, the agreement is poor, although here, too, the cross-correlations appear to be nearly linearly related as the analytic results indicate.

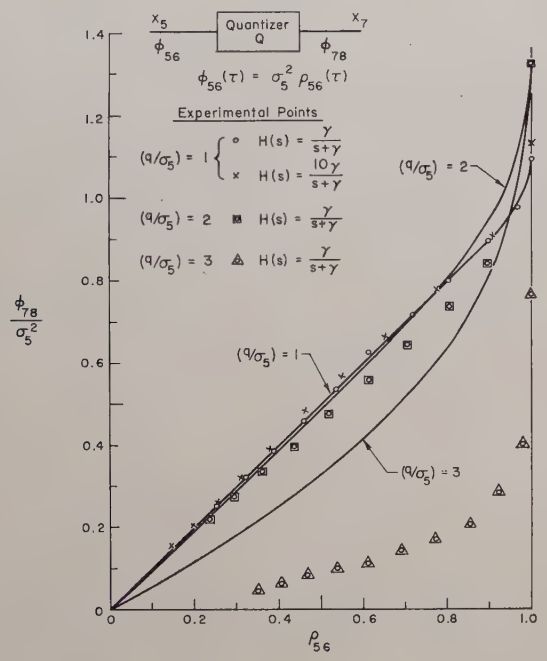


Fig. 5—Comparison of analytical and experimental relation between quantizer autocorrelations.

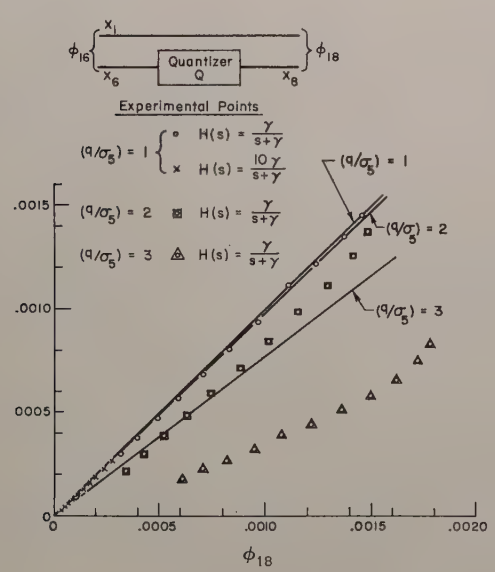


Fig. 6—Comparison of analytical and experimental relation between quantizer cross-correlations.

Beyond this, there are two additional points to note. First, in the one case where experimental data were obtained for both a high-gain system

$$\left[H(s) = \frac{10\gamma}{s+\gamma}\right]$$

and a low-gain system

$$\left[H(s) = \frac{\gamma}{s+\gamma}\right],$$

the agreement between experimental and analytical work appears independent of gain. This implies that the results obtained can be applied fairly widely. The second point concerns the experimental results for $(q/\sigma_5) = 3$. It appears as if the initial slope of the autocorrelation relation of Fig. 5 is equal to the square of the slope of the cross-correlation relation. The former slope is what we have called a; and the latter, K_Q. Thus, it appears that the equality $a = K_Q^2$ still holds for $(q/\sigma_5) = 3$. Furthermore, some of the computer work indicates that the value of K_Q, so determined, is of the right order of magnitude with regard to predicting the error variance of a system with $(q/\sigma_5) \geq 3$.

In Fig. 7 is shown a typical comparison of the analytical with experimental error autocorrelations. The experimental points are based on an $\alpha T = 400$. On this basis, it is felt that there is good agreement between experimental and analytical work. The other experimental data tend to confirm this agreement.

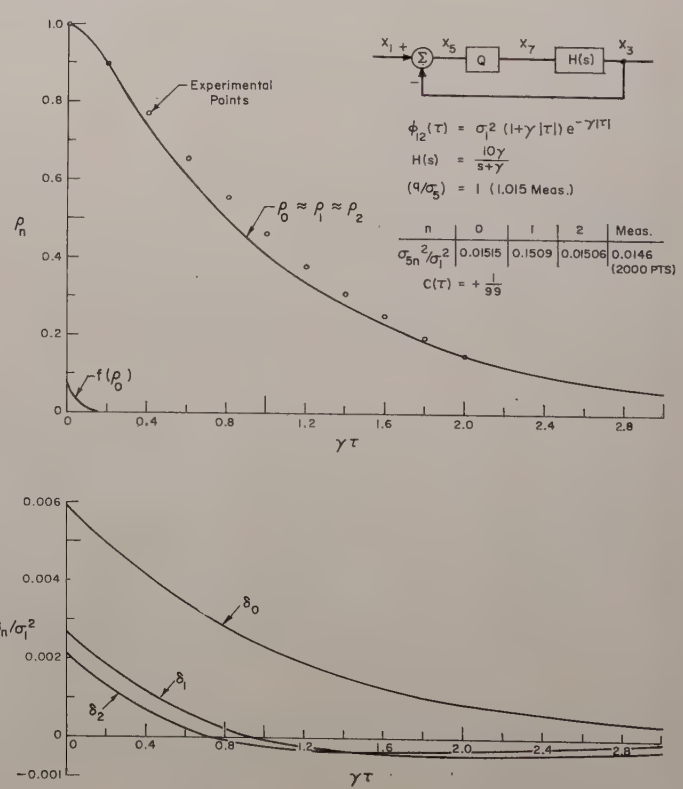


Fig. 7—Experimental and analytical solution to an example: Case 3 $(q/\sigma_5) = 1$.

V. Conclusions

As a result of this study of the effects of quantization, certain conclusions may be drawn. The most important result is that, for the type of quantizer shown in Fig. 1 when $(q/\sigma_5) \leq 2$, the principal effects of quantization, as far as the error autocorrelation is concerned, is a loop-gain reduction introduced by the factor K_Q. The effect of the feedback is to reduce the influence of the nonlinearity introduced by the quantizer. While this situation has been investigated only for a very few examples, it is felt that this conclusion is fairly general. When the value of $(q/\sigma_5) = 3$, the basic assumptions fall down, and the analysis does not give good results. Precisely where, between these values of 2 and 3, the analysis fails has not been investigated. It is in this region that the nonlinearity injected by the quantizer becomes important.

Even though this analysis is not a good approximation in the region of $(q/\sigma_5) > 2$, the general technique of the nonlinear integral equation and its solution by an iterative process does seem to be a good way of solving such problems. If, through a better approximation of the pertinent point distribution or through empirical means, a more accurate approximation of the functional relationships.

$$\phi_{78}(\tau) = g[\phi_{56}(\tau)] \tag{5}$$

and

$$\phi_{18}(\tau) = f[\phi_{16}(\tau)] \tag{6}$$

were available, then this iterative solution technique would provide a useful engineering solution. Indeed, the approach is somewhat more general than implied by its application only to quantizer-type nonlinearities and may be applied profitably to other types of nonlinearities.

With respect to the actual effects of error quantization on the performance of a system, a few fairly general remarks may be made. As indicated above, for $(q/\sigma_5) \leq 2$ the error autocorrelation is essentially equal to that determined by a linear approximation in which the effect of the quantizer is introduced through the gain K_Q. From Fig. 3 it may be seen that this amounts to a gain reduction of less than a few per cent. This is a very small effect considering that, at $(q/\sigma_5) = 2$, the error signal will reach the second level $(x_7 = 2q)$ only 0.26 per cent of the time; under these conditions, the system is essentially a contactor servomechanism. The triviality of the effect under these conditions is most probably due to the effectiveness of the feedback loop in reducing the influence of open-loop nonlinearities on closed-loop behavior.

In addition to these general conclusions, this study pointed up several areas where more investigation is desirable. For one thing, the range of analysis could be extended to larger values of (q/σ_5) by trying to establish a better approximation to the correlation relations of (5) and (6). In addition, ways of modifying the iterative procedure to permit the handling of fixed components with integration must be developed. The convergence problem has sufficiently important and interesting facets to warrant more study. Furthermore, there remains much work to be done in studying other types of quantizers (such as quantizers with nonuniform spacing or with no zero-level output) and effects of combined amplitude quantization and time sampling.

In short, this work represents only a beginning in the understanding of systems in which a discreteness in amplitude and/or time exists.

Appendix I

This appendix presents a brief development of the functional relationship between the autocorrelation of the signal out of the quantizer and that of the input to the quantizer. This development is based on the work of Widrow [4]. The notation used in the main body of this report will be preserved.

A signal $x_5(t)$ with autocorrelation $\phi_{56}(\tau)$ passes through the quantizer and becomes $x_7(t)$. The problem is to find the autocorrelation $\phi_{78}(\tau)$ of the output signal $x_7(t)$. The signal relations and the quantizer gain function are shown in Fig. 8.

Widrow shows that this procedure may be represented as in Fig. 9, where $F_{56}(s_5, s_6; \tau)$ is the characteristic function of $P_{56}(X_5, X_6; \tau)$ and $F_{78}(s_7, s_8; \tau)$ is similarly related to $P_{78}(X_7, X_8; \tau)$. In this representation, P_{56} is treated as a signal (with F_{56} as its transform) which passes through a dynamic element whose frequency response is

$$\frac{\sin s_7 q/2}{s_7 q/2} \frac{\sin s_8 q/2}{s_8 q/2}$$

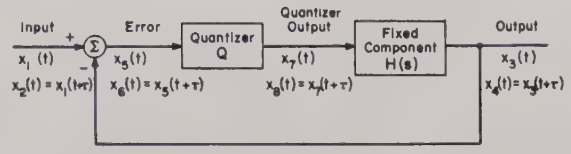
and the resultant signal is then "sampled" (in a two-dimensional manner) with a sampling interval q. This final "signal" is then identical to $P_{78}(X_7, X_8; \tau)$. As a result, the characteristic function of the quantizer output $F_{78}(s_7, s_8; \tau)$ is periodic and consists of a planar array of elements centered at $s_7 = m(2/q), s_8 = n(2/q)$ with each element equal to

$$F_{56}(s_7, s_8) \frac{\sin s_7 q/2}{s_7 q/2} \frac{\sin s_8 q/2}{s_8 q/2}.$$

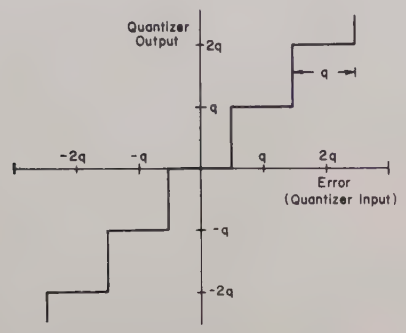
Such a characteristic function is sketched in Fig. 10. From this characteristic function, we may obtain the autocorrelation of the output $\phi_{78}(\tau)$ from that of the input $\phi_{56}(\tau)$ as follows:[2]

$$\begin{aligned} \phi_{78}(\tau) &= \overline{x_7 x_8(\tau)} \\ &= -\left[\frac{\partial^2 F_{78}}{\partial s_7 \partial s_8}\right]_{\substack{s_7=0 \\ s_8=0}}. \end{aligned} \tag{23}$$

[2] See [9], p. 61.



(a)



(b)

Fig. 8—Configuration of feedback system with quantizer. (a) System block diagram. (b) Quantizer gain function.

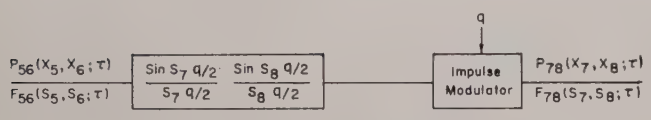


Fig. 9—Operational representation of quantizer effect on input distribution.

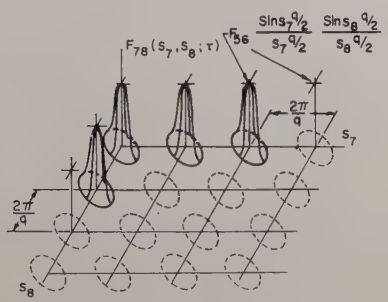


Fig. 10—Joint characteristic function of quantizer output.

For the quantizer

$$F_{78}(s_7, s_8) = \sum_{m,n=-\infty}^{\infty} F_{56}\left(s_7 - m\frac{2\pi}{q}, s_8 - n\frac{2\pi}{q}\right) \cdot \frac{\sin\left(s_7 - m\frac{2\pi}{q}\right)q/2}{\left(s_7 - m\frac{2\pi}{q}\right)q/2} \cdot \frac{\sin\left(s_8 - n\frac{2\pi}{q}\right)q/2}{\left(s_8 - n\frac{2\pi}{q}\right)q/2}. \quad (24)$$

Substituting, taking the derivative, and evaluating at the origin, we arrive at

$$\phi_{78}(\tau) = \sum_{m,n=-\infty}^{\infty} \left\{ \begin{bmatrix} 0 & m = 0,\ n \neq 0 \\ 0 & m \neq 0,\ n = 0 \\ \left(\frac{q}{2\pi}\right)^2 \frac{(-1)^{|m|+|n|}}{mn} & m \neq 0,\ n \neq 0 \end{bmatrix} \cdot F_{56}\left(-m\frac{2\pi}{q}, -n\frac{2\pi}{q}\right) + \begin{bmatrix} 0 & m = 0,\ n \neq 0 \\ \frac{q}{2\pi}\frac{(-1)^{|m|+1}}{m} & m \neq 0,\ n = 0 \\ 0 & m \neq 0,\ n \neq 0 \end{bmatrix} \cdot \frac{\delta F_{56}}{\delta s_8}\left(-m\frac{2\pi}{q}, -n\frac{2\pi}{q}\right) + \begin{bmatrix} \frac{q}{2\pi}\frac{(-1)^{|n|+1}}{n} & m = 0,\ n \neq 0 \\ 0 & m \neq 0,\ n = 0 \\ 0 & m \neq 0,\ n \neq 0 \end{bmatrix} \cdot \frac{\partial F_{56}}{\partial s_7}\left(-m\frac{2\pi}{q}, -n\frac{2\pi}{q}\right) \right\}. \quad (25)$$

At this point, therefore, in order to be able to proceed further, we must assume a characteristic function for the input signal. As we have indicated above, the approximation was made that the characteristic function of the error (input to the quantizer) was essentially that of a Gaussian signal. The basis for this approximation was that, for very small box size (fine quantization), the system approaches linearity and the signals are indeed Gaussian. Proceeding from here, the Gaussian second-order characteristic function is[3]

$$F_{56}(s_5, s_6; \tau) = \exp -\frac{\sigma_5^2}{2}[s_5^2 + s_6^2 + 2\rho_{56}(\tau)s_5 s_6] \quad (26)$$

where

$$\sigma_5^2 = \overline{x_5^2}$$

$$\rho_{56}(\tau) = \frac{1}{\sigma_5^2}\phi_{56}(\tau).$$

Substituting this into (25) above and combining terms result in the following relationship:

[3] *Ibid.*, p. 156.

$$\phi_{78}(\tau) = \phi_{56}(\tau)\left[1 + 4\sum_{n=1}^{\infty}(-1)^n \exp -\frac{1}{2}\left(\frac{2\pi\sigma_5^2}{q}\right)n^2\right] + 2\sigma_5^2\left(\frac{q}{2\pi\sigma_5}\right)^2 \sum_{m,n=1}^{\infty}\frac{(-1)^{m+n}}{mn} \cdot \left\{\exp -\frac{1}{2}\left(\frac{2\pi\sigma_5}{q}\right)^2[m^2+n^2-mn\rho_{56}(\tau)] - \exp -\frac{1}{2}\left(\frac{2\pi\sigma_5^2}{2}\right)[m^2+n^2+mn\rho_{56}(\tau)]\right\}. \quad (27)$$

This relationship is shown plotted in Fig. 2. We may expand this as follows:

$$\phi_{78}(\tau) = a\phi_{56}(\tau) + \sigma_5^2 f[\rho_{56}(\tau)] \quad (28)$$

where

$$a = \left.\frac{d(\phi_{78}/\sigma_5^2)}{d\rho_{56}}\right|_{\rho_{56}=0}$$

$$a = 1 + 4\sum_{n=1}^{\infty}(-1)^n \exp -\frac{1}{2}\left(\frac{2\pi\sigma_5}{q}\right)^2 n^2 + 4\sum_{n=1}^{\infty}(-1)^{m+n}\exp -\left(\frac{2\pi\sigma_5}{q}\right)^2(m^2+n^2). \quad (29)$$

The relations (28) and (29) define $f[\rho_{56}(\tau)]$ as

$$f(\rho_{56}) = \phi_{78}/\sigma_5^2 - a\rho_{56}.$$

Appendix II

Here we consider the situation shown in Fig. 11. In particular, the cross-correlation relationship

$$\phi_{18}(\tau) = f[\phi_{16}(\tau)]$$

is to be determined. Several workers [5], [6], [7] have studied this problem of the correlations of signals passing through a nonlinear, no-memory device and arrived at the conclusion that, for certain classes of signals, the relationship is one of proportionality:

$$\phi_{18}(\tau) = K_Q\phi_{16}(\tau). \quad (30)$$

Fig. 11—Quantizer signals pertinent to cross-correlation relation.

Nutall [5] shows that, if the joint distribution between x_1 and x_5 has the property that he terms "separability," then this proportionality of cross-correlation functions holds. The property of separability is defined as follows:

$$g(X_6;\tau) = \int_{-\infty}^{\infty}(X_1 - \bar{x}_1)P_{16}(X_1, X_6;\tau)dX_1. \quad (31)$$

The joint distribution $P_{16}(X_1, X_6;\tau)$ is separable if $g(X_6;\tau)$ may be broken up into the product of two functions each of which is a function of only one of the variables:

$$g(X_6;\tau) = g_1(X_6)g_2(\tau). \quad (32)$$

Using this it can be shown that

$$\phi_{18}(\tau) = K_Q\phi_{16}(\tau) \quad (30)$$

where

$$K_Q = \frac{\int_{-\infty}^{\infty}dX_6 Q(X_6)g_1(X_6)}{\int_{-\infty}^{\infty}dX_6(X_6 - \bar{x}_6)g_1(X_6)}. \quad (33)$$

This is as far as we may go without specifying the particular joint distribution $P_{16}(X_1, X_6;\tau)$.

Thus far, the only property assumed for P_{16} was that of separability. If we make the further assumption that P_{16} is a Gaussian joint distribution, then we may proceed further since the Gaussian distribution is separable. For a Gaussian distribution[4]

$$P_{16}(X_1, X_6;\tau) = \frac{1}{2\pi\sigma_1\sigma_5\sqrt{1-r^2}}\exp -\frac{1}{2(1-r^2)} \cdot\left[\frac{X_1^2}{\sigma_1^2} + \frac{X_6^2}{\sigma_5^2} - 2r\frac{X_1}{\sigma_1}\frac{X_6}{\sigma_5}\right]$$

where

$$r = r(\tau) = \frac{1}{\sigma_1\sigma_5}\overline{x_1x_6}. \quad (34)$$

Subsituting this into (31) above (setting $\bar{x}_1 = 0$ for zero mean signal), we obtain

$$g(X_6;\tau) = \frac{1}{2\pi\sigma_1\sigma_5\sqrt{1-r^2}}\exp -\frac{1}{2(1-r^2)}\frac{X_6^2}{\sigma_5^2}\int_{-\infty}^{\infty}X_1 \cdot\exp -\frac{1}{2(1-r^2)}\left[\frac{X_1^2}{\sigma_1^2} - 2r\frac{X_1}{\sigma_1}\frac{X_6}{\sigma_5}\right]dX_1. \quad (35)$$

This may be integrated becoming

$$g(X_6;\tau) = \frac{1}{\sqrt{2\pi}}\frac{\sigma_1}{\sigma_5^2}X_6 r\exp\left[-\frac{1}{2}\frac{X_6^2}{\sigma_5^2}\right] = [r(\tau)]\left[\frac{X_6}{\sqrt{2\pi}}\frac{\sigma_1}{\sigma_5^2}\exp\left(-\frac{1}{2}\frac{X_6^2}{\sigma_5^2}\right)\right] \quad (36)$$

$$= g_1(X_6)g_2(\tau). \quad (32)$$

[4] *Ibid.*, p. 78.

From this we find

$$g_1(X_6) = \frac{1}{\sqrt{2\pi}} \frac{\sigma_1}{\sigma_5^2} X_6 \exp\left[-\frac{1}{2}\frac{X_6^2}{\sigma_5^2}\right]$$

$$g_2(\tau) = r(\tau). \tag{37}$$

Substitute these into (33) for K_Q, giving

$$K_Q = \frac{\int_{-\infty}^{\infty} X_6 Q(X_6) \exp\left[-\frac{1}{2}\frac{X_6^2}{\sigma_5^2}\right] dX_6}{\int_{-\infty}^{\infty} X_6^2 \exp\left[-\frac{1}{2}\frac{X_6^2}{\sigma_5^2}\right] dX_6}$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_5^3} \int_{-\infty}^{\infty} X_6 Q(X_6) \exp\left[-\frac{1}{2}\frac{X_6^2}{\sigma_5^2}\right] dX_6. \tag{38}$$

For the quantizer function Q, as shown in Fig. 8, the value of K_Q becomes

$$K_Q = \frac{1}{\sqrt{2\pi}\,\sigma_5^3} \sum_{n=-\infty}^{\infty} nq \int_{(n-1/2)q}^{(n+1/2)q} X_5 \cdot \exp\left[-\frac{1}{2}\frac{X_5^2}{\sigma_5^2}\right] dX_5. \tag{39}$$

This constant K_Q is, implicitly, a function of the normalized box size (q/σ_5). It is shown in Fig. 3.

Appendix III

In Appendix I we sketched the derivation of the relationship between the autocorrelation of the signal into a quantizer and that of the output. This relationship for a Gaussian signal was expanded into

$$\phi_{78}(\tau) = a\phi_{56}(\tau) + \sigma_5^2 f[\rho_{56}(\tau)]. \tag{28}$$

The value of a is shown in (29). Then in Appendix II we showed that the relationship between a cross-correlation at the input of the quantizer to a cross-correlation at the output of the quantizer for Gaussian signals is

$$\phi_{18}(\tau) = K_Q \phi_{16}(\tau). \tag{30}$$

We shall now show that

$$a = K_Q^2 \tag{40}$$

for Gaussian signals. The analysis was suggested by some work of Walker [12].

We consider the system shown in Fig. 11. The cross-correlation at the output is

$$\phi_{18}(\tau) = \int_{-\infty}^{\infty} dX_1 \int_{-\infty}^{\infty} dX_6 X_1 Q(X_6) P_{16}(X_1, X_6; \tau). \tag{41}$$

The joint distribution $P_{16}(X_1, X_6; \tau)$ is assumed to be Gaussian and thus is

$$P_{16}(X_1, X_6; \tau) = \frac{1}{2\pi\sigma_1\sigma_5\sqrt{1-r^2}} \exp -\frac{1}{2(1-r^2)} \cdot \left[\frac{X_1^2}{\sigma_1^2} + \frac{X_6^2}{\sigma_5^2} - 2r\frac{X_1}{\sigma_1}\frac{X_6}{\sigma_5}\right] \tag{42}$$

where

$$\phi_{16} = \sigma_1\sigma_5 r. \tag{43}$$

The functional relationship between x_8 and x_6,

$$x_8 = Q(x_6)$$

across the quantizer is made up of a proportional term and a periodic term which can be expanded into a Fourier series:

$$x_8 = Q(x_6) = x_6 + \sum_{n=0}^{\infty} \frac{(-1)^n}{n} \frac{q}{n} \sin\frac{2\pi n}{q} x_6. \tag{44}$$

Using the relation of (44) in (41), we have

$$\phi_{18}(\tau) = \int_{-\infty}^{\infty} dX_1 \int_{-\infty}^{\infty} dX_6 X_1 \cdot \left[X_6 + \sum_{n=0}^{\infty} \frac{q}{\pi} \sin\frac{2\pi n}{q} X_6\right] P_{16}(X_1, X_6; \tau)$$

$$= \phi_{16}(\tau) + \int_{-\infty}^{\infty} dX_1 \int_{-\infty}^{\infty} dX_6 X_1 \sum_{n=0}^{\infty} \frac{(-1)^n}{n} \frac{q}{\pi} \cdot \left[\sin\frac{2\pi n}{q} X_6\right] P_{16}(X_1, X_6; \tau). \tag{45}$$

Now consider the second term of (45), the integral:

$$I = \sum_{n=0}^{\infty} \frac{(-1)^n}{n} \frac{q}{\pi} \int_{-\infty}^{\infty} dX_6 \sin\frac{2\pi n}{q} X_6 \cdot \int_{-\infty}^{\infty} dX_1 X_1 P_{16}(X_1, X_6; \tau)$$

$$= \sum_{n=0}^{\infty} \frac{(-1)}{n} \frac{q}{\pi} \int_{-\infty}^{\infty} dX_6 \sin\frac{2\pi n}{q} X_6 \cdot \frac{1}{\sqrt{2\pi}} \frac{\sigma_1 r}{\sigma_5^2} X_6 \exp\left[-\frac{1}{2}\frac{X_6^2}{\sigma_5^2}\right] \tag{46}$$

where the integration performed is similar to that in (35) above. Continuing,

$$I = \frac{1}{\sqrt{2\pi}} \frac{\sigma_1 r}{\sigma_5^2} \frac{q}{\pi} \sum_{n=0}^{\infty} \frac{(-1)^n}{n} \int_{-\infty}^{\infty} X_6 \sin\left(\frac{2\pi n}{q} X_6\right) \cdot \exp\left[-\frac{1}{2}\frac{X_6^2}{\sigma_5^2}\right] dX_6. \tag{47}$$

We may integrate this by parts:

$$\int_a^b u\,dv = uv\Big]_a^b - \int_a^b v\,du$$

$$dv = X_6 \exp\left[-\frac{1}{2}\frac{X_6^2}{\sigma_5^2}\right] dX_6$$

$$v = -\sigma_5^2 \exp\left[-\frac{1}{2}\frac{X_6^2}{\sigma_5^2}\right]$$

$$u = \sin\frac{2\pi n}{q} X_6$$

$$du = \frac{2\pi n}{q}\cos\frac{2\pi n}{q} X_6 dX_6. \tag{48}$$

Using these to integrate (42), we obtain

$$I = \frac{1}{\sqrt{2\pi}}\frac{\sigma_1 r}{\sigma_5^2}\frac{q}{\pi}\sum_{n=0}^{\infty}\frac{(-1)^n}{n} \cdot \left\{\left[-\sigma_5^2 \exp\left[-\frac{1}{2}\frac{X_6^2}{\sigma_5^2}\right]\right]_{-\infty}^{0} \sin\frac{2\pi n}{q} X_6\Big]_{-\infty}^{\infty} + \int_{-\infty}^{\infty}\sigma_5^2 \exp\left[-\frac{1}{2}\frac{X_6^2}{\sigma_5^2}\right]\left(\frac{2\pi n}{q}\right)\cos\left(\frac{2\pi n}{q} X_6\right) dX_6\right\}$$

[5]

$$= 2\sigma_1\sigma_5 r \sum_{n=1}^{\infty}(-1)^n \exp\left[-\frac{1}{2}\left(\frac{2\pi\sigma_5}{q}\right)^2 n^2\right]$$

$$I = 2\phi_{16}\sum_{n=0}^{\infty}(-1)^n \exp\left[-\frac{1}{2}\left(\frac{2\pi\sigma_5}{q}\right)^2 n^2\right]. \tag{49}$$

Thus, the total function is, from (45),

$$\phi_{18}(\tau) = \phi_{16}(\tau) + 2\phi_{16}\sum_{n=0}^{\infty}(-1)^n \cdot \exp\left[-\frac{1}{2}\left(\frac{2\pi\sigma_5}{q}\right)^2 n^2\right]$$

$$= \phi_{16}(\tau)\left[1 + 2\sum_{n=0}^{\infty}(-1)^n \cdot \exp\left(-\frac{1}{2}\left(\frac{2\pi\sigma_5}{q}\right)^2 n^2\right)\right]. \tag{50}$$

From the definition of K_Q in (30), we have

$$K_Q = 1 + 2\sum_{n=0}^{\infty}(-1)^n \exp\left[-\frac{1}{2}\left(\frac{2\pi\sigma_5}{q}\right)^2 n^2\right]. \tag{51}$$

[5] This integral is evaluated in R. S. Burington, "Handbook of Mathematical Tables and Formulas," Handbook Publishers, Inc., Sandusky, Ohio, p. 90 (388); 1956.

From this

$$K_Q^2 = \left[1 + 2\sum_{n=0}^{\infty}(-1)^n \exp\left(-\frac{1}{2}\left(\frac{2\pi\sigma_5}{q}\right)^2 n^2\right)\right] \cdot \left[1 + 2\sum_{m=0}^{\infty}(-1)^m \exp\left(-\frac{1}{2}\left(\frac{2\pi\sigma_5}{q}\right)^2 m^2\right)\right]$$

$$= 1 + 4\sum_{n=0}^{\infty}(-1)^n \exp\left[-\frac{1}{2}\left(\frac{2\pi\sigma_5}{q}\right)^2 n^2\right] + 4\sum_{m,n=0}^{\infty}(-1)^{m+n} \exp\left[-\frac{1}{2}\left(\frac{2\pi\sigma_5}{q}\right)^2 (m^2+n^2)\right] \tag{52}$$

which is exactly equal to the value of a in (29). Thus, we have

$$K_Q^2 = a, \qquad \text{Q.E.D.}$$

Appendix IV

This appendix presents an analysis of the errors involved in an experimental measurement of correlation functions. The two sources of error considered here are, first, that due to using periodic samples of the continuous signal of interest instead of the continuous signal itself, and second, that due to using a finite length of data although the time correlation is defined as a limiting process over an infinite length of data. The analysis is based on some work of Davenport, Johnson, and Middleton [10] and of Costas [11].

Suppose we are given a time function $x(t)$ whose autocorrelation is desired. The autocorrelation of this signal may be defined as

$$\phi_{xx}(\tau) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} x(t)x(t+\tau)dt. \tag{53}$$

If, instead of carrying out the averaging process as indicated (which, of course, cannot be carried to the limit) we choose to sample $x(t)$ every t_0 seconds and convert the integration in (53) to a summation; then we sum over a finite number of samples representing a finite length of time T. In carrying out this approximate correlation for various pieces of $x(t)$, we find that the approximate correlation is itself a random variable possessing a mean and a variance. The dependence of these on the sampling period t_0 and the data length T is the object of this analysis.

In making this analysis, we introduce a random variable $Z(t)$ whose approximate average $\dot{Z}$ is obtained by sampling $Z(t)$ every t_0 seconds for a finite time T resulting in N samples. Thus,

$$\dot{Z} = \frac{1}{N}\sum_{n=1}^{N} Z(nt_0) \tag{54}$$

where

$$T = Nt_0. \tag{55}$$

Now $\dot{Z}$ is a new random variable whose mean[6] $\bar{\dot{Z}}$ and whose variance σ_M^2 we desire, since these are measures of our experimental error. The mean of $\dot{Z}$ is

$$\begin{aligned}\bar{\dot{Z}} &= \overline{\frac{1}{N}\sum_{n=1}^{N} Z(nt_0)} \\ &= \frac{1}{N}\sum_{n=1}^{N} \overline{Z(nt_0)} \\ &= \overline{Z(nt_0)}. \end{aligned} \tag{56}$$

Thus, the mean value $\bar{\dot{Z}}$ of the experimental means $\dot{Z}$ is equal to the true mean $\bar{Z}$ of the random variable $Z(t)$. Therefore, our experiment will at least tend to the correct average.

The experimental variance is shown by Costas [11] to be

$$\sigma_M^2 = \frac{\sigma^2}{N} + \frac{1}{N^2}\sum_{k=1}^{N-1} 2(N-k)\phi(kt_0) + m^2\frac{1-N}{N} \tag{57}$$

where

$$\begin{aligned} \sigma^2 &= \overline{Z^2(t)} - m^2, && \text{the true variance of } Z(t) \\ m &= \bar{Z}, && \text{the true average of } Z(t) \\ \phi(\tau) &= \overline{Z(t)Z(t+\tau)}, && \text{the correlation of } Z(t). \end{aligned} \tag{58}$$

Davenport, *et al.*, applied this relation to the determination of a correlation function by making the substitution

$$Z(t, \tau) = x(t)x(t+\tau). \tag{59}$$

The quantities in the equation for σ_M^2 now become

$$\begin{aligned} m(\tau) &= \overline{Z(t)} = \overline{x(t)x(t+\tau)} = \phi_x(\tau) \\ \sigma^2(\tau) &= \overline{Z^2(t)} - m^2 = \overline{x^2(t)x^2(t+\tau)} - m \\ \phi(kt_0) &= \overline{Z(t)Z(t+kt_0)} \\ &= \overline{x(t)x(t+\tau)x(t+kt_0)x(t+kt_0+\tau)}. \end{aligned} \tag{60}$$

At this point we can proceed no further without knowing or approximating some of these just-defined averages. For Gaussian signals, further simplification is possible since they have the special property that the higher-order correlation functions, such as defined above in (60) are related to the simple correlation function:[7]

$$\begin{aligned} &\overline{x(t_1)x(t_2)x(t_1+\tau)x(t_2+\tau)} \\ &\quad = \phi_x^2(t_2-t_1) + \phi_x^2(\tau) \\ &\qquad + \phi_x(t_2-t_1+\tau)\phi_x(t_1-t_2+\tau). \end{aligned} \tag{61}$$

[6] The bar over $\bar{\dot{Z}}$ denotes true average value.
[7] See [9], p. 161.

In order to get at least an approximation to the variance of a sampled, finite-time average, let us make the approximation that the signals with which we are dealing are Gaussian. Therefore, we substitute this relationship in the above equations (60) obtaining

$$\begin{aligned} m(\tau) &= \phi_x(\tau) \\ \sigma^2(\tau) &= \phi_x^2(0) + 2\phi_x^2(\tau) - \phi_x^2(\tau) \\ &= \phi_x^2(0) + \phi_x^2(\tau) \\ \phi(kt_0) &= \phi_x^2(kt_0) + \phi_x^2(\tau) \\ &\quad + \phi_x(kt_0+\tau)\phi_x(\tau-kt_0). \end{aligned} \tag{62}$$

Substituting these into (57) gives

$$\begin{aligned} \sigma_M^2(\tau) = {} & \frac{1}{N}\phi_x^2(0) + \frac{1}{N}\phi_x^2(\tau) \\ & + \frac{2}{N}\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)[\phi_x^2(kt_0) \\ & + \phi_x(\tau+kt_0)\phi_x(\tau-kt_0)] \\ & + \frac{2}{N}\phi_x^2(\tau)\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right) + \frac{1-N}{N}\phi_x^2(\tau). \end{aligned} \tag{63}$$

One of these terms may be summed simply since it is an arithmetic progression:

$$\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right) = \frac{N-1}{2}.$$

This term then cancels the last term giving

$$\begin{aligned} \sigma_M^2 = {} & \frac{1}{N}\phi_x^2(0) + \frac{1}{N}\phi_x^2(\tau) + \frac{2}{N}\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right) \\ & \cdot[\phi_x^2(kt_0) + \phi_x(\tau+kt_0)\phi_x(\tau-kt_0)]. \end{aligned} \tag{64}$$

Some further simplifications result if we make certain approximations. Since we will be treating random variables with zero means, then

$$\phi_x(\tau) \to 0 \quad \text{as} \quad \tau \to \infty.$$

In addition, we will be concerned with long pieces of data so that N will be large and the time shift τ of interest will be small compared to Nt_0. Within these restrictions, we may make the approximation in the summation on the right side of (64) that $\phi_x(kt_0)$ is nonzero only when $k/N \ll 1$. Consequently, we may make the approximation that

$$\begin{aligned} \sigma_M^2 = {} & \frac{1}{N}\phi_x^2(0) + \frac{1}{N}\phi_x^2(\tau) \\ & + \frac{2}{N}\sum_{k=1}^{N-1}[\phi_x^2(kt_0) + \phi_x(\tau+kt_0)\phi_x(\tau-kt_0)]. \end{aligned} \tag{65}$$

This relationship then gives the value of the variance of a finite-time, sampled approximation to an autocorrelation function in terms of the true autocorrelation function,

To obtain some numbers, we assume that the true autocorrelation is of the form

$$\phi_x(\tau) = \sigma_x^2 e^{-\alpha|\tau|}. \tag{66}$$

Substituting in the above

$$\sigma_M^2 = \frac{\sigma_x^4}{N}\left[1 + e^{-2\alpha|\tau|}\right] + \frac{2\sigma_x^4}{N}\sum_{k=1}^{N-1}\left[e^{-2\alpha k t_0} + e^{-\alpha|\tau+kt_0|}e^{-\alpha|\tau-kt_0|}\right]. \tag{67}$$

Considering only values of τ that are positive multiples of t_0, the second term in the summation becomes

$$\sum_{k=1}^{N-1} e^{-\alpha|\tau+kt_0|}e^{-\alpha|\tau-kt_0|} = \sum_{k=1}^{\tau/t_0} e^{-\alpha(\tau+kt_0)}e^{-\alpha(\tau-kt_0)} + \sum_{k=\tau/t_0+1}^{N-1} e^{-\alpha(\tau+kt_0)}e^{+\alpha(\tau-kt_0)} = \tau/t_0 e^{-2\alpha\tau} + \sum_{k=\tau/t_0+1}^{N-1} e^{-2\alpha k t_0}. \tag{68}$$

Substituting back into (67) gives

$$\sigma_M^2(\tau) = \frac{1}{N}\sigma_x^4(1 + e^{-2\alpha\tau}) + \frac{2\sigma_x^4}{N}\sum_{k=1}^{N-1} e^{-2\alpha k t_0} + \frac{2\sigma_x^4}{N}\frac{\tau}{t_0}e^{-2\alpha\tau} + \frac{2\sigma_x^4}{N}\sum_{k=\tau/t_0+1}^{N-1} e^{-2\alpha k t_0}. \tag{69}$$

As assumed above, $\tau \ll Nt_0$; and, consequently, we may neglect the term containing τ/Nt_0. In addition, the summations in (69) represent geometric progressions

$$\sum_{k=p}^{q} e^{-2\alpha k t_0} = \sum_{k=p}^{q} (e^{-2\alpha t_0})^k. \tag{70}$$

The sum of this series is

$$s = b\frac{(1 - r^n)}{(1 - r)}$$

where

$$\begin{aligned} b &= \text{the first term} &&= e^{-2\alpha t_0 p} \\ r &= \text{the ratio of adjacent terms} &&= e^{-2\alpha t_0} \\ n &= \text{the number of terms} &&= q - p + 1. \end{aligned} \tag{71}$$

Therefore,

$$s = e^{-2\alpha t_0 p}\frac{1 - (e^{-2\alpha t_0})^{q-p+1}}{1 - e^{-2\alpha t_0}}. \tag{72}$$

We may now subsitute into (69)

$$\sigma_M^2 = \frac{\sigma_x^4}{N}(1 + e^{-2\alpha\tau}) + \frac{2}{N}\sigma_x^4\left[e^{-2\alpha t_0}\frac{1 - e^{-2\alpha\tau}}{1 - e^{-2\alpha t_0}} + 2e^{-2\alpha(\tau+t_0)}\frac{1 - e^{-2\alpha(Nt_0+\tau+t_0)}}{1 - e^{-2\alpha t_0}}\right]. \tag{73}$$

But $e^{-2\alpha N t_0} \approx 0$ since the time duration of our data was assumed to be large compared to the correlation time constant. We then have

$$\sigma_M^2 = \frac{\sigma_x^4}{N}(1 + e^{-2\alpha\tau}) + \frac{2\sigma_x^4}{N}\left[e^{-2\alpha t_0}\frac{1 - e^{-2\alpha\tau}}{1 - e^{-2\alpha t_0}} + 2\frac{e^{-2\alpha(\tau+t_0)}}{1 - e^{-2\alpha t_0}}\right] = \frac{\sigma_x^4}{N}(1 + e^{-2\alpha\tau})\frac{1 + e^{-2\alpha t_0}}{1 - e^{-2\alpha t_0}}. \tag{74}$$

From (55) we substitute

$$N = T/t_0$$

into (74) obtaining

$$\sigma_M^2 = \frac{\sigma_x^4}{T}(1 + e^{-2\alpha\tau})\frac{t_0(1 + e^{-2\alpha t_0})}{(1 - e^{-2\alpha t_0})}. \tag{75}$$

Within the assumptions and approximations made during the analysis, this equation represents the variance in the approximate correlation function obtained by sampling the random functions every t_0 seconds for a length of time of T seconds. For frequent sampling $\alpha t_0 \ll 1$, (75) becomes approximately

$$\frac{\sigma_x^4}{\alpha T} < \sigma_M^2 < \frac{2\sigma_x^4}{\alpha T}. \tag{76}$$

Bibliography

[1] W. K. Linvill and R. W. Sittler, "Design of Sampled-Data Systems by Extension of Conventional Techniques," Lincoln Lab., Mass. Inst. Tech., Lexington, Rept. R-222; July 3, 1953.

[2] J. G. Truxal, "Control System Synthesis," McGraw-Hill Book Co., Inc., New York, N. Y., ch. 9; 1955.

[3] J. E. Ward, "Digital Cameras—Coders for Spatial Angles," presented at AIEE Fall General Meeting, Chicago, Ill., AIEE Conf. Paper No. 57-1154; October 9, 1957.

[4] B. Widrow, "A study of rough amplitude quantization by means of Nyquist sampling theory," IRE Trans. on Circuit Theory, vol. CT-3, pp. 266–276; December, 1956.

[5] A. H. Nuttall, "Theory and Application of the Separable Class of Random Processes," Sc.D. dissertation, Dept. of Elec. Engrg., Mass. Inst. Tech., Cambridge; May, 1958.

[6] J. Bussgang, "Cross-Correlation Functions of Amplitude-Distorted Gaussian Signals," Res. Lab. of Electronics, Mass. Inst. Tech., Cambridge, Rept. No. 216; 1952.

[7] J. L. Brown, "On a cross-correlation property for stationary random processes," IRE Trans. on Information Theory, vol. IT-3, pp. 28–31; March, 1957.

[8] C. Wagner, "On the solution of the Fredholm integral equation of the second kind by iteration," *J. Math. and Phys.*, vol. 30, pp. 22–30; 1951.

[9] J. H. Laning, Jr. and R. H. Battin, "Random Processes in Automatic Control," McGraw-Hill Book Co., Inc., New York, N. Y.; 1956.

[10] W. B. Davenport, R. A. Johnson, and D. Middleton, "Statistical errors in measurements on random time functions," *J. Appl. Phys.*, vol. 23, pp. 377–388; April, 1952.

[11] J. P. Costas, "Periodic Sampling of Stationary Time Series," Res. Lab. of Electronics, Mass. Inst. Tech., Cambridge, Rept. No. 156; May 16, 1950.

[12] L. Walker, "Optimization of Finite Difference Numerical Integration Procedures," S.M. thesis, Dept. of Elec. Engrg., Mass. Inst. Tech., Cambridge; May, 1957.

[13] Y. Z. Tsypkin, "Elements of the Theory of Computing Automatic Systems," Pre-print Proc. IFAC Congress, Moscow, USSR, June 27–July 7, 1960, Butterworth Scientific Publications, London, England, vol. 2, pp. 997–1003; 1960.

[14] B. Widrow, "Statistical Analysis of Amplitude-Quantized Sampled-Data Systems," presented at AIEE Fall General Meeting, Chicago, Ill., AIEE Trans. Paper No. 60-1240; October 9–14, 1960.

Minimizing Effects of Disturbing Signals Through a Minimum Square-Error Criterion*

MANOEL SOBRAL, JR.†

Summary—One of the reasons for using feedback is the improvement in the rejection of disturbing signals. This improvement can be obtained through an analytical design utilizing as a performance index the integral square-error criterion.[1] In the usual technique[1] the sum of the command signal plus the disturbing signal transferred to the input of the system is used as the input signal. When this is done, one of the two compensating transfer functions (for the particular case of a system with two degrees of freedom) has to be fixed arbitrarily. Then the optimum over-all transfer function, which minimizes the integral of the square of the error between the desired output and the actual one, is calculated and thus the remaining compensator can be obtained. As the technique does not provide a method for determining one of the compensators, and the transferred disturbing signal is a function of this compensator, a required rejection of the disturbing signal may not be satisfied. The purpose of the present paper is to suggest an analytical technique for determining both of the two compensators which have the minimum bandwidth necessary to satisfy a desired over-all transfer function and a required rejection of a disturbing signal. In addition, the technique provides physically realizable compensating transmissions.

I. Statement of the Problem

REFERRING to Fig. 1 (opposite), let $G_f(s)$ be a fixed plant which is subjected to a disturbing signal $U(s)$. Let a certain over-all transfer function $T(s)$ be specified. If the system has only two degrees of freedom, the configuration shown in Fig. 2 can be used without any loss of generality.[2] In this case, the expression for $T(s)$ is

$$T(s) = \frac{G_c(s)G_f(s)}{1 + G_c(s)G_f(s)H(s)} . \tag{1}$$

It can be seen from this expression that $H(s)$ [or $G_c(s)$] can be fixed arbitrarily, and $G_c(s)$ [or $H(s)$] calculated in order to obtain the desired $T(s)$. This freedom of choice of one of the two compensating transfer functions will be used in order to obtain a desired rejection of the disturbing signal.

In order to state the problem, another block diagram which is equivalent to that in Fig. 2 is shown in Fig. 3. Since the transfer function between $R(s)$ and $C(s)$ is specified as $T(s)$, the component of the output $C(s)$ due to $R(s)$ is $C_r(s) = T(s)R(s)$. Insofar as the design of $H(s)$ is concerned, it is sufficient to consider the effects of the disturbing signal alone. Suppressing $R(s)$ in Fig. 3, a new block diagram as in Fig. 4 may be drawn. Note that the input to the new system $H(s)T(s)$ is $U(s)G_{f_2}(s)$, and the output is denoted by $C_u(s)$ which is the part of $C(s)$ due to $U(s)$. It is clear now that if a strong rejection of $U(s)$ is desired, the system transmission $H(s)T(s)$ has to be as close as possible to unity. In this case, $C_u'(s) \approx U(s)G_{f_2}(s)$ and, consequently, $C_u(s)$ will be small.

The problem may now be stated as follows: in order to obtain a good rejection of the disturbing signal $U(s)$, a realizable $H(s)$ has to be chosen in such a way that the transfer function $H(s)T(s)$ is as close as possible to unity. Once $H(s)$ is chosen, $G_c(s)$ can be calculated from (1). In the next paragraph an analytical method to calculate $H(s)$ will be presented.

II. Analytical Design of $H(s)$

From Fig. 4, it is clear that if a total rejection of $U(s)$ is desired, $C_u(s)$ should be zero. In this case $C_u'(s)$ should be equal to $U(s)G_{f_2}(s)$, and the ideal solution for $H(s)$ should be $H(s) = 1/T(s)$. Unfortunately, this solution will lead to a configuration for $H(s)$ which, in general, will have more zeros than poles (as, in general, $T(s)$ has more poles than zeros). Besides, in general, a total rejection of $U(s)$ is not necessary; in most cases, a certain specified degree of rejection is satisfactory. Another important point to be considered is that the bandwidth of $H(s)$ should be minimized for economic reasons.

Considering the reasons above, the problem is to find a certain transfer function $H(s)$, with the minimum bandwidth necessary to satisfy a specified degree of rejection of $U(s)$. The criterion to be used to measure the degree of rejection is the minimum integral square-error criterion. Referring to Fig. 5, this error is defined as

$$I_{yy} = \int_{-\infty}^{+\infty} y_e^2(t)dt, \tag{2}$$

where

$$y_e(t) = C_{du}'(t) - C_u'(t), \tag{3}$$

* Received by the PGAC, November 29, 1960; revised manuscript received, April 20, 1961.

† University of Illinois, Urbana, Ill. On leave of absence from Instituto Tecnologico de Aeronautica, S. Jose dos Campos, S. P., Brazil, where this paper was written.

[1] G. C. Newton, Jr., L. A. Could, and J. F. Kaiser, "Analytical Design of Linear Feedback Controls," John Wiley & Sons, Inc., New York, N. Y.; 1957.

[2] I. M. Horowitz, "Fundamental theory of automatic linear feedback control systems," IRE Trans. on Automatic Control, vol. AC-4, pp. 5–19; December, 1959.

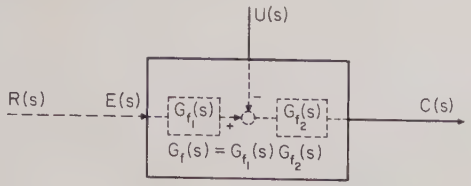


Fig. 1—Block diagram of fixed plant, showing the disturbing signal.

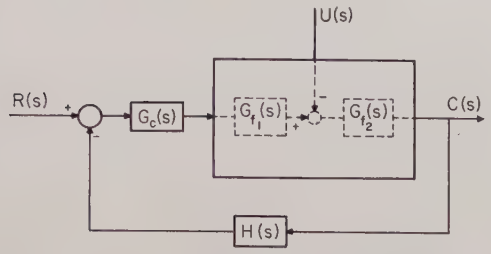


Fig. 2—Proposed configuration for compensating.

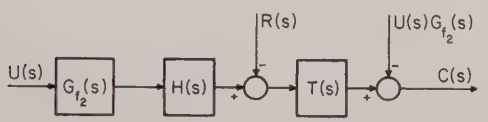


Fig. 3—Modified block diagram for the disturbing signal and command signal.

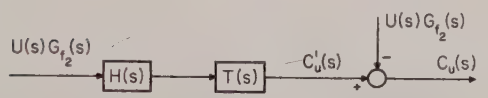


Fig. 4—Block diagram of the equivalent system for the disturbing signal.

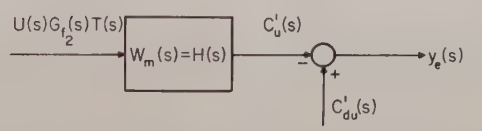


Fig. 5—Block diagram for computing the integral square error and minimizing the bandwidth of H.

and $C_{du}'(t)$ is the ideal response,

$$C_{du}'(t) = U(s)G_{f2}(s). \tag{4}$$

It is clear now that the problem is to find the minimum bandwidth system transfer function $W_m(s) = H(s)T(s)$ which satisfies a constraint in I_{yy} (that is, $I_{yy} \leq I$, where I is a specified upper bound for the integral square error). Since T is fixed, this corresponds to a minimum bandwidth transfer function $H(s)$.

The design of transfer functions for minimum bandwidth satisfying a constraint in the integral square error has been presented already.[3] The solution for $W_m(s)$ is as stated by Newton, *et al.*[4]

$$W_m(s) = \frac{\left[\dfrac{\Gamma(s)}{\Delta^-(s)}\right]_+}{\Delta^+(s)}. \tag{5}$$

Since $W_m(s) = H(s)T(s)$ the following may be written:

$$H(s) = \frac{1}{T(s)} \frac{\left[\dfrac{\Gamma(s)}{\Delta^-(s)}\right]_+}{\Delta^+(s)} \tag{6}$$

where

$$T(s) = kI_{id}(s) \tag{7a}$$

$$\Delta(s) = F(-s)F(s)A(s) + kI_{ii}(s). \tag{7b}$$

[3] Newton, *et al.*, *op. cit.*, ch. 8, pp. 215–243.
[4] Newton, *et al.*, *op. cit.*, ch. 8, p. 227, Eq. 5.4-28.

In the above expressions, $I_{id}(s)$ is the cross-translation function between the input and output signals of the system transfer function $W_m(s)$; $I_{ii}(s)$ is the auto-translation function of the input signal; k is the Lagrange multiplier; $A(s)$ is the autocorrelation function of a noise signal used to minimize the bandwidth of $W_m(s)$; $F(s)$ is the transfer function of an artificial filter used to control the rate of cutoff of $W_m(s)$; $\Delta^+(s)$ is the function which contains as poles and zeros the left-half-plane poles and zeros of $\Delta(s)$; $\Delta^-(s)$ is the function which contains as poles and zeros the right-half-plane poles and zeros of $\Delta(s)$ and $[\Gamma(s)/\Delta^-(s)]_+$ is the Laplace transform of the inverse Fourier transform of $\Gamma(s)/\Delta^-(s)$. In this particular case, the following expressions for I_{id} and I_{ii} can be written

$$I_{ii}(s) = I_{id}(s) = U(-s)G_{f_2}(-s)U(s)G_{f_2}(s). \quad (8a)$$

$A(s)$ is chosen as white noise[3]

$$A(s) = 1. \quad (8b)$$

Since $F(s)$ controls the rate of cutoff of $W_m(s)$, it may be used to assure a number of zeros of $H(s)$ not greater than the number of poles. If $T(s)$ has a rate of cutoff of n db/dec, a convenient choice for $F(s)$ is

$$F(s) = s^m, \qquad m \geq n - 1. \quad (9)$$

One can verify that once $T(s)$ has been chosen with a rate of cutoff at least equal to that of $G_f(s)$, $G_c(s)$ will have a number of poles at least equal to the number of zeros.

It should be mentioned that (6) is restricted to a minimum-phase $T(s)$. This restriction can be removed with the block diagram of Fig. 6, which is equivalent to that of Fig. 5. For this block diagram $H(s)$ can be written as follows:

$$H(s) = \frac{\left[\dfrac{\Gamma(s)}{\Delta^-(s)}\right]_+}{\Delta^+(s)}, \quad (10)$$

and I_{id} and I_{ii} will have the following expressions:

$$I_{id} = U(-s)G_{f_2}(-s)T(-s)U(s)G_{f_2}(s) \quad (11a)$$

$$I_{ii} = U(-s)G_{f_2}(-s)T(-s)U(s)G_{f_2}(s)T(s). \quad (11b)$$

In the next paragraph a simple example will be shown in order to clarify the method.

III. Example

Given the block diagram shown in Fig. 2, where

$$U(s) = \frac{10^{-1}}{s} \quad (12a)$$

$$G_{f_1}(s) = 10 \quad (12b)$$

$$G_{f_2}(s) = \frac{10^{-1}}{s}, \quad (12c)$$

it is desired to calculate $G_c(s)$ and $H(s)$, in order to satisfy

$$T(s) = \frac{10}{s + 10} \quad (13a)$$

and

$$I_{yy}(s) \leq I, \qquad I = \frac{10^{-7}}{2\sqrt{2}}. \quad (13b)$$

In order to employ (6), the following calculations have to be made:

$$U(s)G_{f_2}(s) = \frac{10^{-2}}{s^2} \quad (14a)$$

$$I_{id}(s) = I_{ii}(s) = \frac{10^{-2}}{(-s)^2}\,\frac{10^{-2}}{(s)^2}. \quad (14b)$$

As the rate of cutoff of $T(s)$ is 10 decilog/dec.,

$$F(s) = 1. \quad (15)$$

There results for $H(s)$,

$$H(s) = \frac{s + 10}{10}\,\frac{\left[\dfrac{k\,\dfrac{10^{-2}}{(-s)^2}\,\dfrac{10^{-2}}{(s)^2}}{\left(1 + k\,\dfrac{10^{-2}}{(-s)^2}\,\dfrac{10^{-2}}{s^2}\right)^-}\right]_+}{\left(1 + k\,\dfrac{10^{-2}}{(-s)^2}\,\dfrac{10^{-2}}{(s)^2}\right)^+}. \quad (16)$$

Performing the operations indicated in (16), the following $H(s)$ is obtained:

$$H(s) = \frac{s + 10}{10}\,\frac{k^{1/4}10^{-1}\sqrt{2}\left(s + \dfrac{k^{1/4}10^{-1}}{\sqrt{2}}\right)}{s^2 + k^{1/4}10^{-1}\sqrt{2}\,s + k^{1/2}10^{-2}}. \quad (17)$$

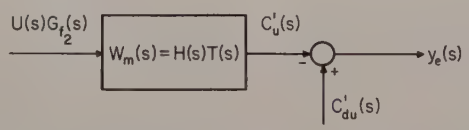


Fig. 6—Block diagram of the equivalent system for nonminimum phase $T(s)$.

Calculating

$$I_{yy} = \int_{-\infty}^{+\infty} y_e^2(t)dt,$$

where

$$y_e(s) = U(s)G_{f_2}(s)[1 - H(s)T(s)],$$

there follows

$$I_{yy} = \frac{10^{-1}}{2\sqrt{2}\,k^{3/4}}. \tag{18}$$

From the given specification for I_{yy}, the value for k is obtained,

$$k = 10^8. \tag{19}$$

There results for $H(s)$

$$H(s) = \frac{s+10}{10}\,\frac{10\sqrt{2}\left(s+\frac{10}{\sqrt{2}}\right)}{s^2+10\sqrt{2}\,s+100}. \tag{20}$$

From (1), $G_c(s)$ is calculated

$$G_c(s) = \frac{10[s^2+10\sqrt{2}\,s+100]}{s(s+10)}. \tag{21}$$

It is interesting to calculate the output of the system due to $U(s)$, that is, $C_u(t)$.

$$C_u(t) = \sqrt{2}\,10^{-2}e^{-5\sqrt{2}t}\cos(\sqrt{2}\,5t - \pi/2). \tag{22}$$

It is worth noticing that the $G_c(s)$ has a pole at the origin which is necessary to assure a finite value for I_{yy}. When the physical configuration for $G_c(s)$ is constructed, this pole can be placed near the origin, and in this case, $C_u(t)$ will have a nonzero steady state.

IV. Conclusions

A new analytical approach to the problem of compensating linear control systems in order to satisfy a given total transfer function and a required rejection of disturbing signals, has been presented. The method utilizes results obtained earlier by other authors. The compensating transfer functions obtained are physically realizable and have the minimum bandwidth necessary to satisfy the specifications.

Appendix

The operations indicated in (16) may be performed as follows:

$$\Delta^-(s) = \left(1 + k\,\frac{10^{-2}}{(-s)^2}\,\frac{10^{-2}}{(s)^2}\right)^- = \left(\frac{s^4 + k\,10^{-4}}{(-s)^2(s)^2}\right)^-$$

$$= \frac{\left[s - k^{1/4}10^{-1}\left(\frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2}\right)\right]\left[s - k^{1/4}10^{-1}\left(\frac{\sqrt{2}}{2} - j\frac{\sqrt{2}}{2}\right)\right]}{(-s)^2} \tag{23}$$

$$\Delta^+(s) = \left(1 + k\,\frac{10^{-2}}{(-s)^2}\,\frac{10^{-2}}{(s)^2}\right)^+$$

$$= \frac{\left[s + k^{1/4}10^{-1}\left(\frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2}\right)\right]\left[s + k^{1/4}10^{-1}\left(\frac{\sqrt{2}}{2} - j\frac{\sqrt{2}}{2}\right)\right]}{s^2} \tag{24}$$

$$\left[\frac{\Gamma}{\Delta^-(s)}\right]_+ = \left[\frac{k\,10^{-4}}{s^2\left[s - k^{1/4}10^{-1}\left(\frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2}\right)\right]\left[s - k^{1/4}10^{-1}\left(\frac{\sqrt{2}}{2} - j\frac{\sqrt{2}}{2}\right)\right]}\right]_+. \tag{25}$$

Calculating the inverse Fourier transform of $\Gamma(s)/\Delta^-(s)$ and obtaining the Laplace transform of the resulting time function, the following expression for $[\Gamma(s)/\Delta^-(s)]_+$ is obtained:

$$\left[\frac{\Gamma(s)}{\Delta^-(s)}\right]_+ = \frac{k^{1/4}10^{-1}\sqrt{2}\left(s + \frac{k^{1/4}10^{-1}}{2}\right)}{s^2}, \tag{26}$$

and (17) can be written by inspection.

I_{yy} is easily obtained using the tables of Newton, *et al.*[1]

Acknowledgment

The author wishes to express his appreciation to Dr. L. V. Boffi for his many valuable and stimulating discussions.

Discussion

Otto J. M. Smith

The solution of Smith[5] can be applied to a block diagram rearrangement of Fig. 4 in which $U(s)$ is the input, and the blocks are in the order $G_{f2}(s)$, $T(s)$, and $H(s)$.

$$X(s) = T(s)G_{f2}(s),\ U(s)$$

is the output of the $T(s)$ block and the input to the $H(s)$ block. $X(s)$ is the unalterable signal, and the $H(s)$ block is to be designed to be the best possible realizable approximation to $1/T(s)$. The solution is

$$H = \frac{1}{\phi_{xx}^{+}}\,\mathfrak{LF}^{-1}\frac{1}{T}\,\phi_{xx}^{+},$$

where

$$\phi_{xx} = T\overline{T}G_{f_2}\overline{G}_{f_2}\phi_{uu}$$

and

$$\phi_{xx}^{+} = TG_{f_2}\phi_{uu}^{+}$$

if T is minimum phase. In order to remove the restriction that T be minimum phase, the methods of Newton, *et al.*,[1] or Smith[5] should be applied to the modified Fig. 4 in which the order of H and T are interchanged, and a noise signal of self-power spectrum $A\overline{F}F$ is added to the output of T, which is the input to H. In this general case, then, (5) will read

$$H = \frac{1}{[F\overline{F}A + kG\overline{G}T\overline{T}\phi_{uu}]^{+}} \cdot \left\{\frac{\overline{T}kG\overline{G}\phi_{uu}}{[F\overline{F}A + kG\overline{G}T\overline{T}\phi_{uu}]^{-}}\right\}_{+}$$

or

$$H(s) = \frac{\left\{\dfrac{T(-s)kI_{ii}(s)}{[F(s)F(-s)A(s) + kT(s)T(-s)I_{ii}(s)]^{-}}\right\}_{+}}{[F(s)F(-s)A(s) + kT(s)T(-s)I_{ii}(s)]^{+}}$$

Let $F(s)F(-s)A(s) = 1.0$; then

$$H = \frac{\left\{\dfrac{k\phi_{xx}}{T[1 + k\phi_{xx}]^{-}}\right\}_{+}}{[1 + k\phi_{xx}]^{+}}$$

where

$$\phi_{xx} = T(s)T(-s)I_{ii}(s).$$

Or, adding an infinitesimally small noise signal α as stated in Smith[5]

$$H = \frac{\left\{\dfrac{\phi_{xx}}{T[\alpha + \phi_{xx}]^{-}}\right\}_{+}}{[\alpha + \phi_{xx}]^{+}} = \frac{1}{[\alpha + \phi_{xx}]^{+}}\,\mathfrak{LF}^{-1}\frac{\phi_{xx}}{T[\alpha + \phi_{xx}]^{-}},$$

where $\alpha < \phi_{xx}$ for all frequencies within the useful bandwidth. Applied to the example in Smith,[6]

$$\phi_{xx} = \frac{10^2}{(10 + s)(10 - s)}\,\frac{10^{-4}}{s^4} = \frac{10^{-2}}{(10^2 - s^2)s^4}.$$

Since T starts to cut off at $s = 10$, consider maximum frequency $s = 20$. At this frequency, $\phi_{xx} \cong 2^{-6} \cdot 10^{-8}$. Let $\alpha = 10^{-10}$.

[5] O. J. M. Smith, "Feedback Control Systems," McGraw-Hill Book Co., Inc., New York, N. Y., (8)–(24) and Figs. 2–8; 1958.

[6] *Ibid.*, p. 227 and Sec. 6–9.

$$H = \frac{\left\{\dfrac{10^{-3}}{(10 - s)s^4\left[\alpha + \dfrac{10^{-2}}{(10^2 - s^2)}\right]^{-}}\right\}_{+}}{\left[\alpha + \dfrac{10^{-2}}{(10^2 - s^2)s^4}\right]^{+}}$$

$$H = \frac{\left\{\dfrac{10^2}{s^2\left[-(s - 21.5)\left(s - 21.5\left(1/2 + j\dfrac{\sqrt{3}}{2}\right)\right)\left(s - 21.5\left(1/2 - j\dfrac{\sqrt{3}}{2}\right)\right)\right]}\right\}_{+}}{\dfrac{10^{-5}\left[(s + 21.5)\left(s + 21.5\left(1/2 + j\dfrac{\sqrt{3}}{2}\right)\right)\left(s + 21.5\left(1/2 - j\dfrac{\sqrt{3}}{2}\right)\right)\right]}{(10 + s)s^2}}$$

$$H = \frac{(10 + s)s^2 \cdot 10^7\left\{\dfrac{10^{-4}\left(1 + \dfrac{s}{10.75}\right)}{s^2}\right\}}{(s + 21.5)\left(s + 21.5\left(1/2 + j\dfrac{\sqrt{3}}{2}\right)\right)\left(s + 21.5\left(1/2 - j\dfrac{\sqrt{3}}{2}\right)\right)}$$

$$H = \frac{93(s+10)(s+10.75)}{(s+21.5)(s+10.75+j10.75\sqrt{3})(s+10.75-j10.75\sqrt{3})}$$

$$H = \frac{(1+s/10)(1+s/10.75)}{\left(1+\frac{s}{21.5}\right)\left(1+\frac{s}{10.75+j10.75\sqrt{3}}\right)\left(1+\frac{s}{10.75-j10.75\sqrt{3}}\right)}.$$

This easily realizable filter can be compared with the case when $\alpha=0$.

$$H = \frac{(10+s)s^2}{10}\left\{\frac{1}{s^2}\right\}_+ = (1+s/10).$$

Obviously the extra poles above are required only to made H easily realizable.

This example is an almost trivial case, since the bandwidths of T and H are the same, and both should be flat out to $s=10$. The approximation in making $H=1.0$ in (20) is no more important than is the selection of I in (13b). A more significant problem would arise if $U(s)=10^{-1}$ in (12a) and $I=10^{-10}$ in (13b).

Manoel Sobral, Jr.

I would like to acknowledge Dr. O. J. M. Smith for his useful remarks. However, an important point should be mentioned. The transfer function $H(s)$ obtained using the general solution for nonminimum-phase $T(s)$ is more complicated than that obtained using (6) for the same set of specifications. The reason for this increased complexity of $H(s)$ can be easily explained. In fact, if $T(s)$ is minimum phase, (6) may be used and, in this case, $H(s)T(s)$ [the system transfer function for the transferred disturbing signal $U(s)G_{f_2}(s)$], may be chosen with a rate of cutoff equal to that of $T(s)$. On the other hand, if the general expression is used, $H(s)T(s)$ will have a rate of cutoff greater than that of $T(s)$ [if $T(s)$ has a cutoff rate of $20n$ db/dec, $H(s)T(s)$ will have a rate of cutoff at least equal to 20 $(n+1)$ db/dec]. Since these two transfer functions $H(s)T(s)$ have to satisfy the same constraint in I_{yy}, the transfer function which has the greater rate of cutoff will necessarily have more bandwidth and/or more gain in certain frequency ranges. Since $T(s)$ is fixed, this improvement has to be achieved through increasing the gain and/or bandwidth of $H(s)$. Concluding, it can be stated that if $T(s)$ is minimum-phase, (6) should be used instead of the general solution in order to obtain a more economical $H(s)$.

Integral Transforms for a Class of Time-Varying Linear Systems*

K. S. NARENDRA†, MEMBER, IRE

Summary—**This paper presents an extension of the transform method to systems having parameters which vary with time. By using the general λ domain approach suggested by Zadeh for the analysis and synthesis of linear time-varying systems, a system function $H(\lambda)$ independent of time may be defined for the linear system. Such a system function has many of the advantages of that obtained for stationary systems using the Laplace transformation. By making $H(\lambda)$ a ratio of polynomials in the complex variable λ the pole-zero synthesis technique used for fixed systems may be applied to the time-varying case as well. Recently, a "building block" for the synthesis of a class of time-varying systems was suggested by Kilmer and Johnson. A similar building block for systems with exponentially varying coefficients is suggested in this paper.**

* Received by the PGAC, November 14, 1960; revised manuscript received, April 28, 1961.

† Harvard University, Cambridge, Mass. Consultant to Minneapolis-Honeywell Regulator Co., Boston Div., Mass.

I. Introduction

LINEAR differential equations with variable coefficients arise from physical systems where some parameter is caused to vary with time because of a process outside the system itself. Time-varying resistances and capacitances in electrical circuits, and varying masses in mechanical systems give rise to such equations. Equations with time-varying coefficients are also known to arise while testing the stability of nonlinear oscillating systems. More recently, in the field of adaptive control, where the parameters of the controller are adjusted to take into account the changes in the process dynamics due to changes in environment, the system differential equations have time-varying coefficients. Since the differential equations are linear,

the tools of linear analysis such as integral transforms and superposition integrals may be used while attempting their solution. A general approach to the analysis of time-varying systems is very difficult and at the present stage of development in the field only the solution of a relatively simple class of problems arising in practical applications is attempted. This report reviews and extends the use of generalized integral transforms for the analysis and synthesis of certain classes of linear time-varying systems.

II. The Integral Transform Method

The use of integral transforms to solve dynamical problems in physics and engineering has received considerable attention in the last half century. The technique has been made rigorous and several specific transform pairs have been developed. This, in turn, has led to a generalization of transform methods. The use of the Laplace transform method for linear lumped parameter systems is well known. The transform converts a linear differential equation with constant coefficients into an algebraic equation in the transform variable s. The advantages of working in the transform domain rather than in the time domain are quite well known in the fields of control systems and network synthesis. In the case of time-varying linear systems where the integro-differential equations have time-varying coefficients, suitable transform techniques may be developed to obtain the same advantages. Throughout the report, the correspondence between the methods used for the time-varying and constant coefficient cases, as well as the basic differences in viewpoint, will be emphasized.

The general differential equation relating a single input and a single output variable of a linear time-varying system may be expressed as

$$\left\{a_n(t)\frac{d^n}{dt^n} + a_{n-1}(t)\frac{d^{n-1}}{dt^{n-1}} + \cdots + a_0(t)\right\}\theta_0(t) = \left\{b_m(t)\frac{d^m}{dt^m} + \cdots + b_0(t)\right\}\theta_i(t) \quad (1)$$

where $\theta_0(t)$ and $\theta_i(t)$ represent the output and input variables (Fig. 1).

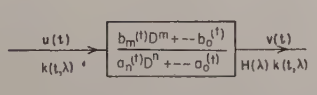


Fig. 1—Linear time-varying system.

To obtain the solution of the linear differential equation (1) by the integral transform method, we define a new function $U(\lambda)$ of the complex variable $\lambda = \xi + i\eta$ by means of:

$$u(t) = \int_c k(t, \lambda)U(\lambda)d\lambda, \quad (2)$$

where the transform kernel $k(t, \lambda)$ and the path of integration c are to be suitably determined. For purposes of analysis and synthesis of signal transmission systems, $k(t, \lambda)$ may be considered as an "elementary" component signal into which $u(t)$ may be resolved. Then $U(\lambda)$ is the spectral function of $u(t)$ relative to $k(t, \lambda)$. It is assumed that the kernel $k(t, \lambda)$ has an inverse $k^{-1}(\lambda, t)$, and that the relation between the two functions has been defined by Zadeh [5] as

$$\int_c k(t, \lambda)k^{-1}(\lambda, t')d\lambda = \delta(t - t'). \quad (3)$$

Using (2) and (3) the spectral function $U(\lambda)$ may be written in terms of $k^{-1}(\lambda, t)$ as

$$U(\lambda) = \int_{-\infty}^{\infty} u(t)k^{-1}(\lambda, t)dt. \quad (4)$$

In the case of systems with constant coefficients, the analysis is carried out in the frequency domain using the Laplace transform where $e^{st}/2\pi i$, the transform kernel, represents the elementary signals, and c is the Bromwich-Wagner contour.

If the response of the system to an input $k(t, \lambda)$ is $K(t, \lambda)$, the response $v(t)$ to an input function $u(t)$ by the principle of superposition is given by

$$v(t) = \int_c K(t, \lambda)U(\lambda)d\lambda. \quad (5)$$

If $K(t, \lambda)$ is expressed as

$$K(t, \lambda) = H(\lambda)k(t, \lambda), \quad (6)$$

from (5) we have

$$v(t) = \int_c H(\lambda)U(\lambda)k(t, \lambda)d\lambda, \quad (7)$$

or $H(\lambda)\,U(\lambda)$ is the λ transform of $v(t)$, *i.e.*, $V(\lambda)$. Thus, the λ transform of the output function is seen to be the product of the transform of the input and $H(\lambda)$ which is consequently defined as the "system function" (Fig. 2). If the transform pair defined by (2) and (4) *is unique*,

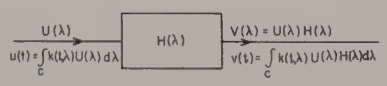


Fig. 2—System function of time-varying system.

the system function $H(\lambda)$ defines the system completely and analysis and synthesis of the system may be attempted in the λ domain. The problem is thus reduced to the determination of $k^{-1}(\lambda, t)$, $k(t, \lambda)$, and the contour c. If for a given system $H(\lambda)$ can be expressed as a ratio of polynomials in λ, the pole-zero synthesis technique used quite extensively in the control systems and network fields may be extended to the case of time-vary-

ing systems. In a constant coefficient system, an input function e^{st} yields a response $W(s)e^{st}$ (where $W(s)$ is the transfer function of the system) and hence, the Laplace transform approach is satisfactory. In the case of time-varying systems which do not have a general transient response of the form

$$\sum_{1}^{n} A_i e^{-\sigma_i t} \sin (w_i t + \Phi_i)$$

such an input would yield a system function $H(s, t)$ which is a function of time. The "elementary" signal $k(t, \lambda)$ that is used is consequently of the form of a homogeneous solution of the system differential equation.

Determination of $k(t, \lambda)$

The differential equation governing the system is given by (1) and if $\theta_i(t) = k(t, \lambda)$ and $\theta_0(t) = H(\lambda)k(t, \lambda)$, we have

$$\left\{a_n(t)\frac{d^n}{dt^n} + \cdots + a_0(t)\right\} H(\lambda)k(t, \lambda) = \left\{b_m(t)\frac{d^m}{dt^m} + \cdots + b_0(t)\right\} k(t, \lambda) \quad m < n \tag{8}$$

which may be reduced to the form

$$\left\{c_n(t)\frac{d^n}{dt^n} + \cdots + c_0(t, \lambda)\right\} k(t, \lambda) = 0 \tag{9}$$

where

$$c_j(t) = a_j(t) - \left[\frac{1}{H(\lambda)}\right] b_j(t) \qquad 0 < j < m$$
$$= a_j(t) \qquad m < j < n.$$

Hence, the success of the transform method depends on our ability to solve (9).

First-Order Equation

The simplest type of time-varying differential equation that has a closed-form solution is

$$k'(t, \lambda) + \left[\frac{1}{H(\lambda)}\right] f(t)k(t, \lambda) = 0 \qquad ' = \frac{\delta}{\delta t} \tag{10}$$

and $k(t, \lambda)$ is given by

$$k(t, \lambda) = \exp\left[\frac{-1}{H(\lambda)}\int^{t} f(x)dx\right]. \tag{11}$$

Thus, any $H(\lambda)$ yields a corresponding $k(t, \lambda)$. $H(\lambda)$ is chosen to make the determination of the inverse kernel simple. In a recent paper [7], Kilmer and Johnson have shown that if $H(\lambda) = 1/\lambda$ is the system function of a variable gain integrator, the transform pair, (2) and (4), may be justified with the Laplace transform pair after a suitable change of variables.

Second-Order System

The determination of $k(t, \lambda)$ for a second-order system is not possible for arbitrary time-varying coefficients. For example, if (1) is of the form

$$\left\{a(t)\frac{d^2}{dt^2} + b(t)\frac{d}{dt}\right\} \theta_0(t) = \theta_i(t), \tag{12}$$

(9) may be expressed as

$$\left\{a(t)\frac{d^2}{dt^2} + b(t)\frac{d}{dt} - \frac{1}{H(\lambda)}\right\} k(t, \lambda) = 0. \tag{13}$$

For definite values of $a(t)$ and $b(t)$, (13) reduces to a standard form such as the Bessel, Legendre, or Euler equation, and in such cases the system function $H(\lambda)$ may be suitably chosen to make $k(t, \lambda)$ simple. The block diagram representation of these equations is shown in Fig. 3(a) and (b).

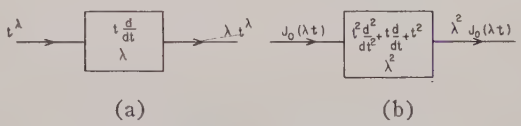


Fig. 3—Block diagram representation of differential equations. (a) Euler equation. (b) Bessel equation.

Once $k(t, \lambda)$ is chosen, if the integral in (2) can be identified with the transform integral of an existing transform pair, the inverse kernel $k^{-1}(\lambda, t)$ and the contour c may be determined by inspection. If such an identification is not possible, $k^{-1}(\lambda, t)$ and c have to be determined using the relation given by (3).

Determination of $k^{-1}(\lambda, t)$ for a Second-Order System

If the differential equation governing the system is given by (12), $k(t, \lambda)$ is obtained by solving (13). If for a specific system function $H(\lambda)$, $k_1(\lambda, t)$, and $k_2(\lambda, t)$ are the two independent solutions of (13) and $C_1(\lambda)k_1(\lambda, t)$ is chosen as the kernel $k(t, \lambda)$, the inverse kernel $k^{-1}(\lambda, t)$ is given by

$$k^{-1}(\lambda, t) = k_2(\lambda, t) \exp\left[\int \frac{b - a'}{a} dt\right] \qquad a' = \frac{\partial a}{\partial t} \tag{14}$$

(See Appendix I). The contour c is chosen so that

$$\int_c k_1(t, \lambda)k_2(\lambda, t')C_1(\lambda) \exp\left[\int^{t'} \frac{b - a'}{a} dt\right] d\lambda = \delta(t - t'). \tag{15}$$

We have chosen $k^{-1}(\lambda, t)$ in order to make the integral in (4) converge.

Example

If $a(t)=1$, $b(t)=1/t$, and $H(\lambda)$ is chosen as $1/\lambda^2$, (13) is of the form

$$k''(\lambda, t) + \frac{1}{t}k'(\lambda, t) - \lambda^2 k(\lambda, t) = 0 \qquad '' = \frac{\partial^2}{\partial t^2}$$

$$k_2(\lambda, t) = K_0(\lambda t) \qquad k_2(\lambda, t) = I_0(\lambda t) \quad (16)$$

where K_0 and I_0 are modified Bessel functions of the first and second kinds. If $\lambda I_0(\lambda t)$ is chosen as the transform kernel, the inverse kernel $k^{-1}(t, \lambda)$ is given by

$$k^{-1}(t, \lambda) = tK_0(\lambda t).$$

We then have the transform pair

$$\int_{-\infty}^{\infty} tK_0(\lambda t)u(t)dt = U(\lambda)$$

and

$$\frac{1}{\pi j}\int_{\beta-j\infty}^{\beta+j\infty} \lambda I_0(\lambda t)U(\lambda)d\lambda = u(t). \quad (17)$$

The contour is given by

$$\int_c \lambda I_0(\lambda t)t'k_0(\lambda t')d\lambda = \delta(t - t')$$

and is, hence, the Bessel contour.

III. Analysis and Synthesis of Systems in the λ Domain

When the transform pair specified by (2) and (4) is unique, $H(\lambda)$ completely defines the system and corresponds to the transfer function in the linear constant coefficient case. In Section II it was shown that the λ transform of the output function is the product of the input transform and the system function. Hence, the system function of two systems (governed by the same differential equation) in series is the product of the system functions, *i.e.*, $H^2(\lambda)$. By a general cascading of the same system defined by $H(\lambda)$ (in series, feedback, and parallel), system functions which are ratios of polynomials in $H(\lambda)$ may be obtained. If $H(\lambda)=\lambda^{\pm r}$ where r is an integer, the system functions so developed would be a ratio of polynomials in λ. For such classes of systems the well-known pole-zero analysis and synthesis techniques used in the linear constant coefficient case may be extended to the λ domain. The singularities are interpreted in terms of the time response of the system.

The kernel functions e^{st} and e^{-st} being the same for all linear stationary systems, then any two systems with transfer functions $H_1(s)$ and $H_2(s)$ in series will yield a transfer function $H_1(s)H_2(s)$. For the time-variable case the transform pair developed applies only to the particular differential equation under consideration. Consequently, only one building block is available for the cascading process as shown in Fig. 4. Two or more building blocks may be used if the transform pair developed simultaneously applies to two or more differential equations. This implies that $k(t, \lambda)$, $k^{-1}(\lambda, t)$ and the contour c are common to the various systems.

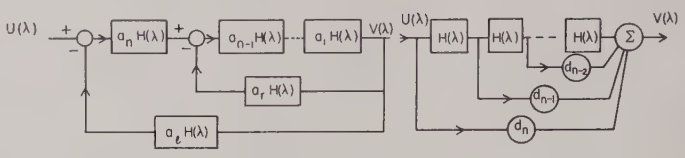


Fig. 4—General cascading of system function of linear time-varying system.

Some Useful Relations

In the analysis carried out so far, the impulse response of the system has not been considered. The response of the system to an arbitrary input $u(t)$ may be obtained by considering (7) where $u(t)$ is resolved into its component elementary signals $k(t, \lambda)$ (*i.e.*, considering the λ domain) or by a superposition integral involving the impulse response. The delta function has a uniform frequency spectrum when the Fourier transform is considered and consequently the system function of a constant coefficient system is also the transform of its impulse response. However, the impulse function does not have a uniform spectrum with respect to $k(t, \lambda)$ and consequently the relation between the impulse response and the system function $H(\lambda)$ is not straightforward. Let $w(t, \zeta)$ be the response of the system to a unit impulse at $t=\zeta$; the response of the system to an arbitrary input $u(t)$ is given by

$$v(t) = \int_{-\infty}^{\infty} w(t, \zeta)u(\zeta)d\zeta \quad (18)$$

$$V(\lambda) = U(\lambda)H(\lambda) = \int_{-\infty}^{\infty} k^{-1}(\lambda, t)v(t)dt$$

$$= \int_{-\infty}^{\infty} k^{-1}(\lambda, t)\left[\int_{-\infty}^{\infty} w(t, \zeta)u(\zeta)d\zeta\right]dt \quad (19)$$

or changing the order of integration

$$= \int_{-\infty}^{\infty} u(\zeta)d\zeta \int_{-\infty}^{\infty} k^{-1}(\lambda, t)w(t, \zeta)dt$$

$$= \int_{-\infty}^{\infty} W(\lambda, \zeta)u(\zeta)d\zeta \quad (20)$$

where $W(\lambda, \zeta)$ is the λ transform of the impulse response. Thus

$$H(\lambda) = \frac{1}{U(\lambda)}\int_{-\infty}^{\infty} W(\lambda, \zeta)u(\zeta)d\zeta. \quad (21)$$

From (4) the transform of $\delta(t-\zeta)$ is $k^{-1}(\lambda, \zeta)$. Thus

$$W(\lambda, \zeta) = H(\lambda)k^{-1}(\lambda, \zeta). \quad (22)$$

λ Domain Representation of Differential Equations

When the differential equation relating the input and the output of a system is given by

$$\left\{a_n(t)\frac{d^n}{dt^n} + a_{n-1}(t)\frac{d^{n-1}}{dt^{n-1}} + \cdots + a_0(t)\right\} v(t) = u(t) \quad (23)$$

and the system function by $H(\lambda)$,

$$V(\lambda) = U(\lambda)H(\lambda) \quad (24)$$

or

$$\int_{-\infty}^{\infty}\left\{a_n(t)\frac{d^n}{dt^n} + \cdots + a_0(t)\right\} v(t)k^{-1}(\lambda, t)dt = \frac{V(\lambda)}{H(\lambda)}. \quad (25)$$

Thus $k^{-1}(\lambda, t)$, the inverse kernel, may also be considered as one which converts the differential equation in $v(t)$ to $V(\lambda)/H(\lambda)$ in the λ domain [6]. We may, therefore, represent standard equations like the Bessel and Euler equations by their corresponding kernels $k^{-1}(\lambda, t)$ and system functions $H(\lambda)$. Block diagram representations are shown in Fig. 3. Higher-order equations may then be generated by a general cascading of these basic building blocks. The fact that transform pairs already exist in such cases simplifies the problem of determining $k^{-1}(\lambda, t)$ and the contour c.

Bessel Equation

In the example previously considered $H(\lambda)$ is chosen as $-\lambda^2$, the transform kernels are given by

$$k(t, \lambda) = J_0\left(\frac{1}{\lambda}t\right) \qquad k^{-1}(\lambda, t) = tY_0\left(\frac{1}{\lambda}t\right).$$

The output $v(t)$ to any input $u(t)$ is given by

$$\frac{1}{\pi j}\int_{\beta-j\infty}^{\beta+j\infty}\frac{U(\lambda)}{\lambda^2}J_0\left(\frac{t}{\lambda}\right)d\lambda$$

where $\beta < \sigma$

$$\int_0^{\infty} e^{\beta t}v(t)dt \quad (26)$$

converges.

Euler Equation

When the differential equation relating input and output is the Euler equation

$$\left[t^2\frac{d^2}{dt} + t\frac{d}{dt}\right]\theta_0 = \theta_i(t) \quad (27)$$

$$k_1(t, \lambda) = t^{-\lambda} \qquad k_2(t, \lambda) = t^{\lambda}$$

$$a(t) = t^2 \qquad b(t) = t \qquad \exp\left[\int\frac{b - a'}{a}dt\right] = \frac{1}{t}.$$

Therefore, $k(\lambda, t) = t^{-\lambda}$ and $k^{-1}(\lambda, t) = t^{\lambda-1}$.

The Mellin transform pair is given by

$$U(\lambda) = \int_0^{\infty} u(t)t^{\lambda-1}dt \quad (28)$$

$$u(t) = \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} U(\lambda)t^{-\lambda}d\lambda \quad (29)$$

and, hence, the contour is the same as the Bromwich-Wagner contour.

For the transform pair and contour specified, $H(\lambda) = +1/\lambda^2$. The differential equation may also be represented as

$$\left\{t\frac{d}{dt}\right\}\left\{t\frac{d}{dt}\right\}\theta_0 = \theta_i, \quad (30)$$

so that $t(d/dt)\theta_0 = \theta_i$ may be considered as the basic building block with a system function $1/\lambda$.

By a general cascading of the basic building block equations of the form

$$\left[t^n\frac{d^n}{dt^n} + A_1t^{n-1}\frac{d^{n-1}}{dt^{n-1}} + \cdots + A_n\right]\theta_0(t) = \theta_i(t) \quad (31)$$

may be solved in the λ domain.

Systems with Exponentially-Varying Coefficients

In many practical time-varying systems, the coefficients can be approximated by an exponential function of the form $(A + Be^{-at})$ over a finite period of time. Electrical networks with large time constants subject to pulse inputs and mechanical systems with varying mass are known to give rise to such equations.

An equation of the type

$$\frac{d^2\theta}{dt^2} + 2Ae^{\pm t}\frac{d\theta}{dt} + Be^{\pm t}\theta = 0 \quad (32)$$

is known to arise while considering the dynamics of a missile decelerating through a variable atmosphere [8] when the density variation is assumed to be exponential with respect to altitude. The first step in the transform method is the determination of the transform kernel $k(t, \lambda)$ by solving the differential equation relating the input and output functions. Certain types of first-order and second-order equations with exponential coefficients can be solved analytically and in such cases building blocks may be set up in the transform domain as mentioned previously.

First-Order Equation

A special case of the equation treated by Kilmer and Johnson [7] is one in which $f(t) = e^{\mp at}$. The differential equation is of the form (Fig. 5)

$$e^{\pm at}\frac{d\theta_0}{dt} = \theta_i(t) \quad (33)$$

and the system function and transform kernels are given by

$$H(\lambda) = \frac{1}{\lambda} \qquad k(t, \lambda) = e^{\lambda T} \qquad k^{-1}(\lambda, t) = e^{-\lambda T} \quad (34)$$

where

$$T = \int_0^t e^{\mp ax} dx.$$

By making the above transformation and identifying the transform pair with the Laplace transform pair, the contour c is identified with the Bromwich-Wagner contour.

By a general cascading of the block shown in Fig. 4 an over-all system function of the form

$$Y_{0v}(\lambda) = \frac{k}{\lambda^n + d_1\lambda^{n-1} + \cdots + d_n} \quad (35)$$

can be obtained.

This, however, corresponds to a restricted class of differential equations with time-varying coefficients. For example, the general second-order equation that can be solved using this approach is of the form

$$\left\{\frac{d^2}{dt^2} + (Ae^{\pm at} \pm a)\frac{d}{dt} + Be^{\pm 2at}\right\}\theta_0(t) = \theta_i(t). \quad (36)$$

In order to generate more general systems new building blocks in the λ domain are required.

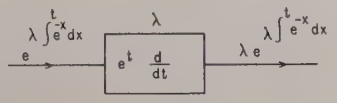


Fig. 5—Building block for linear systems with exponentially-varying coefficients.

Second-Order Equation

The second-order equation

$$\ddot{\theta} + 2A\dot{\theta} + (C^2e^{\pm 2t} - B^2)\theta = 0 \quad (37)$$

may be transformed into

$$\left\{T^2\frac{d^2}{dT^2} + T(1 \pm 2A)\frac{d}{dT} + (C^2T^2 - B^2)\right\}\theta = 0 \quad (38)$$

by the transformation $e^{\pm t} = T$.

Eq. (38) is the Bessel equation whose solution is given by

$$\theta(T) = [C_1J_p(CT) + C_2J_{-p}(CT)]T^{\mp A} \qquad p = \sqrt{A^2 + B^2}. \quad (39)$$

When p is zero or a positive integer, the second independent solution is given by

$$C_2Y_p(CT)T^{\mp A}$$

where Y_p is the Bessel function of the second-kind and order p.

Many other special cases in which the system equation with exponentially-varying coefficients can be reduced to a standard form using the above transformation are shown in Appendix II. Since a solution in closed form exists in these cases, the transform method may be attempted.

Let the input and output be related by the equation

$$e^{\mp 2t}[\ddot{\theta}_0 + 2A\dot{\theta}_0] = \theta_i(t). \quad (40)$$

The transform kernel $k(t, \lambda)$ satisfies the equation

$$k''(t, \lambda) + 2Ak'(t, \lambda) - \frac{e^{\pm 2t}}{H(\lambda)}k(t, \lambda) = 0. \quad (41)$$

If $H(\lambda) = -1/\lambda^2$, (41) is of the form of (37), and the solutions are given by

$$k(t, \lambda) = C_1T^{\mp A}J_A(\lambda T) \qquad T = e^{\pm t}. \quad (42)$$

If $A = 0$

$$k(t, \lambda) = C_1J_0(\lambda T).$$

The Hankel transform pair is given by

$$u(T) = \int_0^\infty \lambda J_0(\lambda T)U(\lambda)d\lambda \quad (43)$$

$$U(\lambda) = \int_0^\infty TJ_0(\lambda T)u(T)dT \quad (44)$$

choosing $C_1 = \lambda$ and identifying $k(t, \lambda)$ with the kernel of the Hankel transform

$$k^{-1}(\lambda, t) = TJ_0(\lambda T) \quad (45)$$

since the Hankel transform pair is unique, $H(\lambda) = -1/\lambda^2$ completely defines the system, and is represented in Fig. 6.

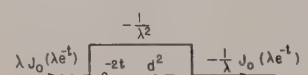


Fig. 6—Building block for linear systems with exponentially-varying coefficients.

The response of the system to an arbitrary input $\theta_i(t) = u(t)$ is given by

$$v(T) = -\int_0^\infty \frac{U(\lambda)}{\lambda}J_0(\lambda T)d\lambda \quad (46)$$

where

$$U(\lambda) = \int_0^\infty u(T)TJ_0(\lambda T)dT$$

and

$$V(\lambda) = \int_0^\infty TJ_0(\lambda T)v(T)dT. \quad (47)$$

A general cascading of this single block yields system functions of the form

$$\frac{k}{\lambda^{2n} + A_1\lambda^{2n-2} + A_2\lambda^{2n-4} + \cdots + A_n} \tag{48}$$

which again generates a restricted class of linear differential equations. The most general second-order equation is of the form

$$e^{\pm 2t}[\ddot{\theta}_0(t) + A\dot{\theta}_0(t)] = \theta_i(t) \tag{49}$$

where A is an arbitrary constant.

IV. Conclusion

The integral transform method transforms a linear differential equation with time-varying coefficients into an algebraic equation in the λ domain. By developing suitable building blocks in the λ domain and using a general cascading procedure, classes of differential equations can be generated for which the particular transforms developed would apply. The single integrator and the double integrator with exponentially-varying gain which have been treated are specific examples. If the system function of the building block is a ratio of polynomials in λ, the cascading procedure yields an over-all system function which can be analyzed by the well-known pole-zero synthesis technique. The fact that the transform pair developed applies to a single differential equation implies that only a single building block can be used in the cascading procedure. The class of differential equations that can be solved in this manner is very limited. However, the development of integral transform pairs which apply simultaneously to two or more differential equations will, it is hoped, remove this serious restriction.

Appendix I

Relation Between the Transform Kernels $k(t, \lambda)$ and $k^{-1}(\lambda, t)$ for a Second-Order System

Let the input and output of a time-varying system be related by the differential equation

$$\left[a(t)\frac{d^2}{dt^2} + b(t)\frac{d}{dt}\right]v(t) = u(t). \tag{50}$$

The kernel $k(t, \lambda)$ suggested in this paper satisfies the equation

$$a(t)k''(t, \lambda) + b(t)k'(t, \lambda) - \frac{1}{H(\lambda)}k(t, \lambda) = 0 \tag{51}$$

where $H(\lambda)$ is the system function and is chosen to make $k(t, \lambda)$ simple. The inverse transform $k^{-1}(\lambda, t)$ is defined by

$$U(\lambda) = \int_{-\infty}^{\infty} k^{-1}(\lambda, t)u(t)dt \tag{52}$$

substituting for $u(t)$

$$U(\lambda) = \int_{-\infty}^{\infty}\left[a(t)\frac{d^2v}{dt^2} + b(t)\frac{du}{dt}\right]k^{-1}(\lambda, t)dt. \tag{53}$$

But $V(\lambda) = H(\lambda)U(\lambda)$ by definition, so $k^{-1}(\lambda, t)$ is the transform kernel which transforms a differential equation in $v(t)$ to the form $V(\lambda)/H(\lambda)$. It has been shown by Aseltine [6] that under zero-initial conditions the inverse kernel can be represented as

$$k^{-1}(\lambda, t) = g(t)R(\lambda, t) \tag{54}$$

where

$$g(t) = \exp\left[\int \frac{b - a'}{a}dt\right]$$

and $R(\lambda, t)$ satisfies the equation

$$a(t)R''(\lambda, t) + b(t)R'(\lambda, t) - \frac{1}{H(\lambda)}R(\lambda, t) = 0. \tag{55}$$

Hence, $k(t, \lambda)$ and $R(\lambda, t)$ both satisfy the same differential equation (55). Let $k_1(\lambda, t)$ and $k_2(\lambda, t)$ be two linearly independent solutions of the equation. Any linear combination of the two solutions is a solution of the differential equation. We must choose $R(\lambda, t)$ so that

$$\int_{-\infty}^{\infty} k^{-1}(\lambda, t)v(t)dt$$

converges. Considering signals $v(t)$ which are zero for $t<0$

$$\int_{0}^{\infty} k^{-1}(\lambda, t)v(t)dt \tag{56}$$

must converge, and consequently, $k_2(\lambda, t)$ is chosen as $R(\lambda, t)$, if

$$k_2(\lambda, t) \to 0 \qquad t \to \infty. \tag{57}$$

If $k_1(\lambda, t)$, is chosen so that

$$\int_{c} k_1(\lambda, t)k_2(\lambda, t')\exp\left[\int^{t'}\frac{b - a'}{a}dt\right]d\lambda = \delta(t - t') \tag{58}$$

$k_1(\lambda, t)$ is the needed transform kernel. The transform pair developed depends on the initial choice of $k_2(\lambda, t)$.

For the stationary case, when $a(t) = a$ and $b(t) = b$

$$a\frac{d^2}{dt^2} + b\frac{d}{dt} - \left(\frac{1}{H(\lambda)}\right)R(\lambda, t) = 0$$

choosing

$$H(\lambda) = \frac{1}{a\lambda^2 + b\lambda} \qquad k_1(\lambda, t) = e^{+\lambda t}k_2(\lambda, t)$$

$$= \exp\left[-\left(\frac{b}{a} + \lambda\right)t\right].$$

Hence,

$$R(\lambda, t) = k_2(\lambda, t) = \exp\left[-\left(\frac{b}{a} + \lambda\right)t\right].$$

The transform kernel is

$$k^{-1}(\lambda, t) = e^{+\lambda t}$$

and the inverse kernel is

$$k^{-1}(\lambda, t) = \exp\left[-\left(\frac{b}{a} + \lambda\right)t\right] \cdot \exp\left[\int \frac{b}{a}\,dt\right] = e^{-\lambda t}$$

$$\int_c k(t, \lambda)k^{-1}(\lambda, t)d\lambda = \int_c e^{+(t-t')\lambda}d\lambda = \delta(t - t') \quad \text{for} \quad t' \text{ positive.}$$

Appendix II

Second-Order Equations with Exponentially-Varying Coefficients: Special Cases

It was shown in the paper that

$$\ddot{\theta} + 2A\dot{\theta} + (c^2e^{\pm 2t} - B^2)\theta = 0$$

may be transformed into

$$\left[T^2\frac{d^2}{dT^2} + (1 \pm 2A)T\frac{d}{dT} + (c^2T^2 - B^2)\right]\theta = 0$$

using the transformation $e^{\pm t} = T$, and the solutions of this equation are given by

$$[c_1J_p(cT) + c_2J_p(cT)]T^{\mp A} \qquad p = \sqrt{A^2 + B^2}.$$

Other notable special cases where the equations reduce to known forms by a similar transformation for special values of the coefficients are mentioned below:

1) $$(1 - e^{\pm at})\ddot{\theta} + B\dot{\theta} + Ae^{\pm at}\theta = 0 \qquad (59)$$

using the transformation $e^{\pm at} = T$ reduces to the hypergeometric equation

$$T(1 - T)\frac{d^2\theta}{dT^2} + \left(1 \pm \frac{B}{a} - T\right)\frac{d\theta}{dT} + \frac{A}{a^2}\theta = 0. \qquad (60)$$

2) The equation

$$\frac{d^2\theta}{dt^2} - \frac{a^2}{4}(1 \pm 2e^{\pm at})\frac{d\theta}{dt} + na^2e^{\pm t}\theta = 0, \qquad (61)$$

if n is a positive integer, is satisfied by the Hermite polynomial of degree n, while

$$(1 - e^{\pm at})\frac{d^2\theta}{dt^2} \pm \frac{a^2}{4}(1 + e^{\pm at})\frac{d\theta}{dt} + n^2a^2e^{\pm at}\theta = 0 \qquad (62)$$

is satisfied by the Tchebycheff polynomial $T_n(e^{\pm at/2})$.

$$(1 - e^{\pm at})\ddot{\theta} \pm \frac{a^2}{4}(2 + e^{\pm at})\theta + a^2p(p + 1)e^{\pm at} = 0 \qquad (63)$$

is reduced to the Legendre equation.

4) The equation

$$\ddot{\theta} + (A + e^{-t})\dot{\theta} - ae\theta^{-t} = 0 \qquad (64)$$

is transformed to the "confluent hypergeometric function" of Kummer, and for the special case when $A = 1$ and $a = -n$ where n is a positive integer, one solution is the nth Laguerre polynomial.

Acknowledgment

The author takes this opportunity to express his thanks to T. Baker for many helpful discussions.

References

[1] L. A. Zadeh, "Frequency analysis of variable networks," Proc. IRE, vol. 38, pp. 291–299; March, 1950.
[2] ——, "Circuit analysis of linear varying parameter networks," *J. Appl. Phys.*, vol. 21, pp. 1171–1177; November, 1950.
[3] ——, "The determination of the impulsive response of variable networks," *J. Appl. Phys.*, vol. 21, pp. 642–645; July, 1950.
[4] ——, "Theory of filtering signals in the 'λ' domain," *Bull. Am. Math. Soc.*, vol. 57, p. 278; July, 1951 (Abstract).
[5] ——, "General input output relations for linear networks," Proc. IRE (Correspondence), vol. 40, p. 103; January, 1952.
[6] J. A. Aseltine, "A transform method for linear time varying systems," *J. Appl. Phys.*, vol. 25, pp. 761–764; June, 1954.
[7] G. W. Johnson and F. G. Kilmer, "Integral Transforms for Algebraic Analysis and Design of a Class of Linear Variable and Adaptive Control Systems," presented at Joint Automatic Control Conference, Mass. Inst. Tech., Cambridge; September 7–9, 1960.
[8] H. J. Allen, "Motion of a Ballistic Missile Angularly Misaligned with the Flight Path upon Entering the Atmosphere and Its Effect upon Aerodynamic Heating, Aerodynamic Loads, and Miss Distance," NACA Tech. Note No. 4048; 1957.

Signal Stabilization of Self-Oscillating Systems*

R. OLDENBURGER† AND T. NAKADA‡

Summary—The hunt (self-oscillations) of a physical system may often be removed by the introduction of an appropriate stabilizing signal which changes the open loop gain in a nonlinear manner. More generally, the performance of nonlinear systems in many cases may be improved by the introduction of extra signals. The theory of signal stabilization developed here extends the earlier work by Oldenburger and Liu involving an equivalent gain concept. It is shown that with the aid of the Fourier series the designer can determine the periodic signal to be inserted at one point in a loop to yield a desired stabilizing input to a nonlinear element in the loop. The use of sinusoidal and triangular inputs to a limiter are compared. An example where a limiter is the only nonlinearity is employed to illustrate the theory. The approach developed here explains experimental results previously reported by Oldenburger.

I. Introduction

IN 1957, Oldenburger reported the insertion of a signal to remove the hunt or self-oscillations of unstable nonlinear systems, and gave some experimental results.[1] In 1959, Oldenburger and Liu published a theoretical explanation for this phenomenon which agreed with the experiment.[2] The explanation differs from that given by Minorsky regarding the asynchronous quenching of systems described by nonlinear differential equations.[3] The use of extra signals to improve the performance of nonlinear elements has been known for a long time. Dither has been employed to minimize the effects of dry friction, hysteresis and other phenomena. MacColl in 1945 reported the use of a sinusoidal signal to linearize a relay.[4] Lozier, in 1950, discussed carrier controlled relay servos.[5] The present paper gives some results of a study by Oldenburger and associates on the effect of extra signals on the input-output characteristics of nonlinear elements. By the use of such signals these characteristics may often be transformed so that a given nonlinearity behaves as if it were quite different. In many cases the introduction of an extra signal is less expensive than actually replacing the nonlinear element by one with the characteristics obtainable from the original by the use of the extra signal. Also undesired signals, often called "noise," are always present in physical systems, modifying the input-output characteristics of nonlinearities so that they behave differently from what might be otherwise expected. The effect of undesired extra signals on the performance of nonlinear systems becomes particularly important when, as in the control of space missiles and satellites, one is concerned with threshold signals.

In the Oldenburger and Liu paper, the problem of determining a sinusoidal stabilizing signal for a given system with a limiter as the only nonlinear element was solved for a basic closed loop system. In the present paper, this theory is extended to rather general closed loop systems with one nonlinearity and to the determination of a stabilizing signal at the input to a system, that will yield a given periodic input to the nonlinear element. This is illustrated with a triangular input to the nonlinearity. The use of triangular and sinusoidal stabilizing signals are compared. These results should assist the control designer in choosing a stabilizing signal to remove the hunt, or self-oscillation, of a given nonlinear system or improve its performance.

II. Equivalent Gain

This paper is restricted to nonlinear components described by functions. That is, if x is the input and y the output of the nonlinear component, there is a function $f(x)$ so that

$$y = f(x). \tag{1}$$

The characteristic of the nonlinear component is thus assumed to be time-independent. The theory may be extended to components that do not satisfy this requirement. Nonlinear components described by functions will be called functional nonlinearities. Consider such a nonlinear element NL as shown in Fig. 1, where

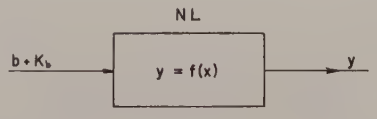


Fig. 1—Nonlinear element NL with input $x=b+K_b$.

the input x is composed of two components b and K_b, so that

$$x = b + K_b \tag{2}$$

for a constant K_b and a stabilizing signal b.

Let y_{av} denote the average output of NL. Assume that the average of the stabilizing signal b is zero, so that K_b is the average of x. It is assumed that y_{av} is zero

* Received by the PGAC, December 12, 1960.
† Dept. of Mechanical Engrg., Purdue University, Lafayette, Ind.

[1] R. Oldenburger, "Signal stabilization of a control system," *Trans. Am. Soc. Mech. Engrg.*, vol. 69, pp. 1869–1872; August, 1957. Translations into Japanese and Russian were published in the *Japanese J. Automatic Control* and the *Avtomat. i Telemekh.*, respectively.

[2] R. Oldenburger and C. C. Liu, "Signal stabilization of a control system," *Trans. AIEE*, vol. 78 (*Commun. and Electronics*, no. 59-219), pp. 96–100; May, 1959.

[3] N. Minorsky, "On asynchronous action," *J. Franklin Inst.*, vol. 259, pp. 209–219; March, 1955.

[4] L. A. MacColl, "Fundamental Theory of Servomechanisms," D. Van Nostrand Co., Inc., New York, N. Y., pp. 78–87; 1945.

[5] J. C. Lozier, "Carrier-controlled relay servos," *Elec. Engrg.*, vol. 69, pp. 1052–1056; December, 1950.

when $K_b=0$. The equivalent gain g_b of NL is by definition given by

$$g_b = \lim_{K_b \to 0} \frac{y_{av}}{K_b} . \tag{3}$$

The above averages are for $t=-\infty$ to $t=+\infty$. In the general definition of the equivalent gain g_b the restriction $K_b \to 0$ is dropped, whence g_b is a function of K_b. However, the case where $K_b \to 0$ will suffice in this paper.

The limiter with characteristic shown in Fig. 2 is a commonly occurring nonlinearity. Here the output reaches the limit B when the input attains the value a. Let the stabilizing input signal b to the limiter be the sine wave $B_1 \sin \omega t$ with amplitude B_1 and radian frequency ω, where t denotes time. The equivalent gain g_b of this limiter is given for this stabilizing signal b by

$$g_b = \frac{2B}{\pi a} \operatorname{Sin}^{-1} \frac{a}{B_1} \tag{4}$$

as proved in Appendix I. A plot of g_b vs B_1 may be made as shown in Fig. 3 for $B=a$.

In treating the stability of a physical system with functional nonlinearities, each nonlinearity is replaced by a simple amplifier or attenuator with the gain g_b, thus linearizing the system. Linear design techniques may then be applied to achieve the characteristic roots that will ensure satisfactory stability, and performance for small distrubances. For large disturbances the equivalent gain y_{av}/K_b with $K_b \neq 0$ may be employed, but we shall not be concerned with such disturbances here.

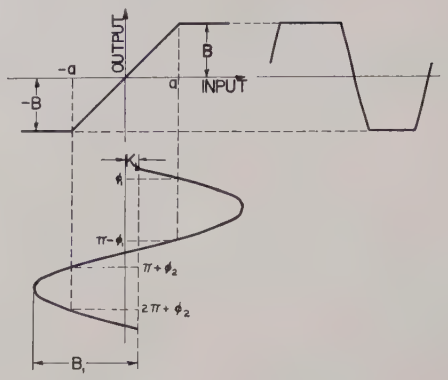


Fig. 2—Characteristic of limiter.

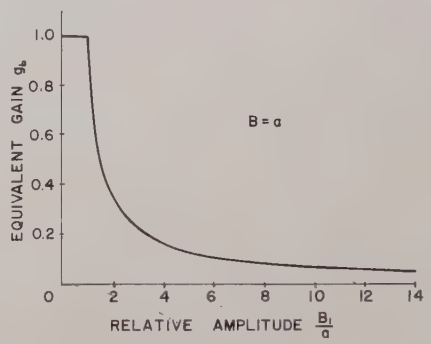


Fig. 3—Equivalent gain of limiter.

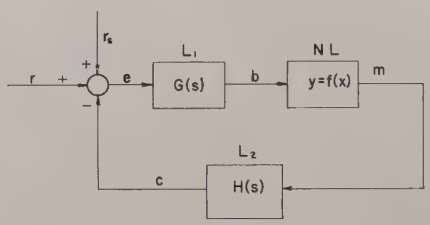


Fig. 4—Nonlinear closed loop system.

III. Determination of a Periodic Stabilizing Signal

This paper concerns the closed loop system of Fig. 4 where a linear element L_1 follows the input to the system, a nonlinear element NL follows L_1 and this is in turn followed by a linear element L_2. The output of L_2 is subtracted from the input $r+r_s$, where r_s is the signal introduced to stabilize the system, and r is the normal input to the system. With $r=0$ the presence of r_s results in an input stabilizing signal b to the element NL. The determination of a periodic signal r_s, given the signal b, will be considered. The output of the element NL due to the input b to NL will be denoted by m.

In design practice, the procedure would be as follows. For computational purposes the element NL is replaced by a proportional element with gain g_b, so that the output of the element is g_b times the input. By linear design the desired value of g_b is determined. From the relation between the equivalent gain g_b and the nature of the stabilizing signal b (a relation assumed to have been determined previously), the precise signal b that will give the gain g_b is determined. The designer must now choose r_s in the absence of the signal r (*i.e.* with $r=0$) so as to obtain b at the input to the nonlinear element NL. A procedure for doing this will be outlined. The procedure will be illustrated later for the case of a triangular wave b.

If the given stabilizing signal b is periodic, it may be assumed that b is given by the Fourier series

$$b = \sum_{n=1}^{\infty} b_n \sin (n\omega t + \phi_n) \tag{5}$$

for a radian frequency ω, phase angles ϕ_n, and constants b_n. The output of the nonlinear element may be assumed to be the periodic wave m, where

$$m = \sum_{n=1}^{\infty} m_n \sin (n\omega t + \psi_n) \tag{6}$$

for constants $\{m_n\}$ and phase angles $\{\psi_n\}$. The quantities m_n and ψ_n are mathematically determined by formula 5 for b and the function $f(b)$, where

$$m = f(b) \tag{7}$$

and can be computed by the use of the standard in-

tegral formulae for coefficients in Fourier series. The output c of the linear element L_2 may be written as

$$c = \sum_{n=1}^{\infty} c_n \sin (n\omega t + \lambda_n) \tag{8}$$

for constants $\{c_n\}$ and $\{\lambda_n\}$. These constants can be directly computed from the transfer function $H(s)$ of L_2 and the Fourier series m, which represents the input to L_2. Each component $m_n \sin (n\omega t+\psi_n)$ can be written in complex form and multiplied by the transfer function to yield the complex form of the component $c_n \sin (n\omega t+\lambda_n)$ of the output c.

The input stabilizing signal r_s may also be expanded in a Fourier series,

$$r_s = \sum_{n=1}^{\infty} R_n \sin (n\omega t + \mu_n) \tag{9}$$

for constants $\{R_n\}$ and $\{\mu_n\}$. The constants $\{R_n\}$ and $\{\mu_n\}$ are to be determined.

The input to the linear element L_1 will be denoted by e, where (with $r=0$)

$$e = r_s - c. \tag{10}$$

From the transfer function $G(s)$ of L_1 and the complex representations of the terms in the Fourier expansion of the output b of L_1, the constants e_n and ν_n in the Fourier series

$$e = \sum_{n=1}^{\infty} e_n \sin (n\omega t + \nu_n) \tag{11}$$

are determined. It follows from relations (8)–(11) that

$$R_n \sin (n\omega t + \mu_n) = e_n \sin (n\omega t + \nu_n) + c_n \sin (n\omega t + \lambda_n). \tag{12}$$

Expansion of the trigonometric functions yields

$$\begin{aligned} R_n \cos \mu_n &= e_n \cos \nu_n + c_n \cos \lambda_n \\ R_n \sin \mu_n &= e_n \sin \nu_n + c_n \sin \lambda_n. \end{aligned} \tag{13}$$

These relations may now be solved for R_n and μ_n.

Thus, by the steps outlined above, a periodic stabilizing signal r_s may be determined to yield a desired equivalent gain g_b for the nonlinear element NL. Similarly, the input stabilizing signal r_s may be related to the equivalent gains of nonlinear elements in loops with two or more nonlinear elements separated by linear components.

IV. Equivalent Gain for Triangular Stabilizing Signal

The example of Fig. 5 given in a paper by Liu and Oldenburger[2] will be considered. This is one of the simplest physical cases to which signal stabilization applies. This example arose in industry, where K_1+sK_2 was the transfer function of the controller, m' the speed of the servo output of the controller, $1/s^2(\tau s+1)^2$ the transfer function of the system controlled and c the controlled variable. Here s is the standard Laplace variable. The prime on m in Fig. 5 denotes the derivative dm/dt.

It shall be assumed that the stabilizing signal at the input b to the limiter in Fig. 5 is the triangular wave shown in Fig. 6 with period 2π and amplitude B_1 of which one period is shown in the illustration. The input signal r_s to the loop that will yield this wave at the input to the limiter will be determined. Here and in the rest of the paper all harmonics are retained. The function $f(\theta)$ representing this triangular wave is given by

$$f(\theta) = \frac{8B_1}{\pi^2}\left(\sin\theta - \frac{\sin 3\theta}{3^2} + \frac{\sin 5\theta}{5^2} - \cdots\right). \tag{14}$$

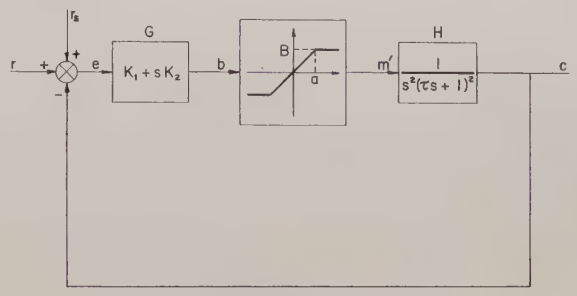


Fig. 5—Nonlinear system.

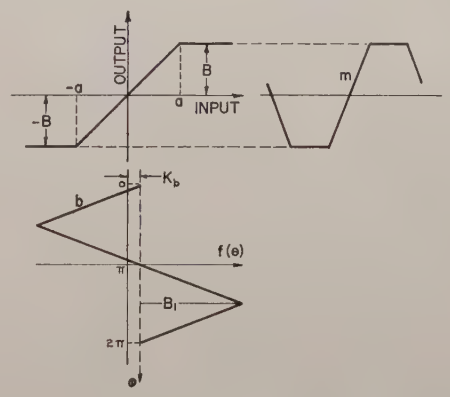


Fig. 6—Limiter input-output when input is biased triangular wave.

Suppose now that the input b to the limiter of Fig. 5 is the triangular wave $f(\theta)$ biased by K_b, *i.e.*,

$$b = K_b + f(\theta). \tag{15}$$

It is assumed that the input b reaches saturation of the limiter at both the top and bottom of the wave b, as shown in Fig. 6.

The output m of the limiter is now

$$m = \frac{B}{B_1}K_b + \frac{8B_1}{\pi^2}\left(\frac{B}{a}\right) \cdot \left[\sum_{n=2,4\cdots}^{\infty} \frac{1}{n^2} \sin\frac{n\pi}{2}\left(\frac{a}{B_1}\right) \sin\frac{n\pi}{2}\left(\frac{K_b}{B_1}\right) \cdot \cos n\theta + \sum_{n=1,3,5\cdots}^{\infty} \frac{1}{n^2} \sin\frac{n\pi}{2}\left(\frac{a}{B_1}\right) \sin n\theta\right]. \tag{16}$$

The average value of the limiter output is given by

$$m_{av} = \left(\frac{B}{B_1}\right) K_b. \tag{17}$$

The *equivalent gain* of the limiter is now g_b where

$$g_b = \lim_{K_b \to 0} \frac{m_{av}}{K_b} = \frac{B}{B_1}. \tag{18}$$

V. Comparison of Sinusoidal and Triangular Waves

Let a sine wave input to the limiter of Fig. 5 be given by

$$A \sin \omega t. \tag{19}$$

Let m_s be the amplitude of the fundamental sinusoidal component of the output of the limiter. Now[2]

$$m_s = A\left(\frac{B}{a}\right)\frac{2}{\pi}\left\{\sin^{-1}\frac{a}{A} + \left(\frac{a}{A}\right)\left[1 - \left(\frac{a}{A}\right)^2\right]^{1/2}\right\}. \tag{20}$$

If (a/A) is small enough, we may approximate m_s as in

$$m_s \approx \frac{2B}{\pi}\left\{2 - \frac{1}{3}\left(\frac{a}{A}\right)^2\right\}. \tag{21}$$

From (17) when the bias K_b is zero, the amplitude of the fundamental component m_t of the output signal of the limiter due to the input triangular wave of Fig. 6 is given by

$$m_t = 8\left(\frac{B_1}{\pi^2}\right)\left(\frac{B}{a}\right)\sin\frac{\pi}{2}\left(\frac{a}{B_1}\right). \tag{22}$$

For (a/B_1), sufficiently small, we may approximate m_t as

$$m_t \approx B\left\{\frac{4}{\pi} - \frac{\pi}{6}\left(\frac{a}{B_1}\right)^2\right\}. \tag{23}$$

Now a triangular wave equivalent to the sine wave $A \sin \omega t$ as one having the same frequency and slope at crossings of the x axis as a sine wave, illustrated in Fig. 7, is defined. The maximum height B_{1e} of the equivalent triangular wave is then given by

$$B_{1e} = \frac{\pi}{2} A. \tag{24}$$

Substituting in (21) yields

$$m_s \approx B\left\{\frac{4}{\pi} - \frac{\pi}{6}\left(\frac{a}{B_{1e}}\right)^2\right\}. \tag{25}$$

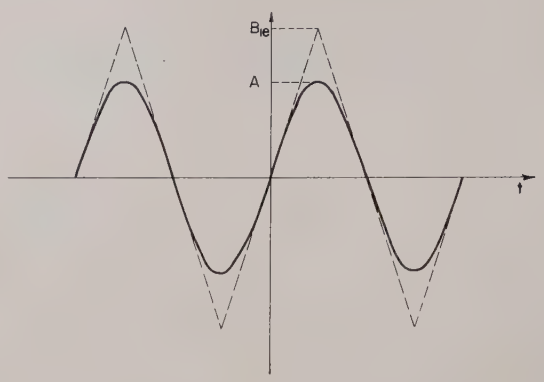


Fig. 7—Equivalent sine and triangular waves.

It may be noted that (23) and (25) are identical if B_1 and B_{1e} are identified as well as m_t and m_s. It is concluded that the output of the limiter to either a sinusoidal input or an equivalent triangular input is the same up to the second order of (a/B_{1e}). This indicates that there is not much difference between using a sinusoidal stabilizing input to the limiter and a triangular input.

VI. Relation Between Loop Input Signal and Triangular Stabilizing Input to Limiter

It will be shown how a stabilizing signal r_s for the system of Fig. 5 will yield a given triangular stabilizing signal b at the input to the limiter. The triangular input to the limiter may be written as

$$b = I_m \sum_{n=0}^{\infty} U_{2n+1} e^{j(2n+1)\omega t}, \tag{26}$$

where

$$U_{2n+1} = \frac{8B_1}{\pi^2}\frac{(-1)^n}{(2n+1)^2}. \tag{27}$$

Here I_m denotes "the imaginary part of," j is $-1^{1/2}$, and ω is the radian frequency of the wave. The limiter output is m', where

$$m' = I_m \sum_{n=0}^{\infty} M_{2n+1} e^{j(2n+1)\omega t} \tag{28}$$

and

$$M_{2n+1} = \frac{8B_1}{\pi^2}\left(\frac{B}{a}\right)\frac{1}{(2n+1)^2} \cdot \sin\left\{(2n+1)\left(\frac{\pi}{2}\right)\left(\frac{a}{B_1}\right)\right\}. \tag{29}$$

The stabilizing signal r_s required is the imaginary part of R_s, where

$$R_s = \sum_{n=0}^{\infty} A_{2n+1} e^{j\{(2n+1)\omega t + \theta_{2n+1}\}}. \tag{30}$$

Here (A_{2n+1}) and (θ_{2n+1}) are to be determined. We introduce G_{2n+1}, β_{2n+1}, H_{2n+1} and α_{2n+1} given by

$$\begin{aligned} G_{2n+1} &= \left| (K_1 + sK_2)_{s=(2n+1)j\omega} \right| \\ &= \left\{ K_1^2 + (2n+1)^2\omega^2 K_2^2 \right\}^{1/2} \\ &= \text{Mag}\left\{ G(\overline{2n+1}\cdot j\omega) \right\}, \end{aligned} \tag{31}$$

where Mag denotes "Magnitude of."

$$\beta_{2n+1} = \text{Arg}\left\{ G(\overline{2n+1}\cdot j\omega) \right\} = \tan^{-1}(2n+1)\frac{\omega K_2}{K_1} \tag{32}$$

where Arg means "argument of." Thus,

$$G(\overline{2n+1}\cdot j\omega) = G_{2n+1}e^{j\beta_{2n+1}}. \tag{33}$$

Also

$$\begin{aligned} H_{2n+1} &= \left| \left[\frac{1}{s^2(\tau s+1)^2} \right]_{s=(2n+1)j\omega} \right| \\ &= \frac{1}{(2n+1)^2\omega^2\left\{ (2n+1)^2\omega^2\tau^2 + 1 \right\}} \\ &= \text{Mag}\left\{ H(\overline{2n+1}\cdot j\omega) \right\} \end{aligned} \tag{34}$$

$$\alpha_{2n+1} = \text{Tan}^{-1}(2n+1)\omega\tau. \tag{35}$$

Thus,

$$H(\overline{2n+1}\cdot j\omega) = -H_{2n+1}e^{-2j\alpha_{2n+1}}. \tag{36}$$

The complex form C of the output c of the system of Fig. 5 is given by

$$C = HM \tag{37}$$

for the complex form M of m'. By formula (28)

$$M = \sum_{n=0}^{\infty} M_{2n+1}e^{j(2n+1)\omega t}. \tag{38}$$

By (34)–(38), there results

$$C = -\sum_{n=0}^{\infty} H_{2n+1}M_{2n+1}e^{j\{(2n+1)\omega t - 2\alpha_{2n+1}\}}. \tag{39}$$

The complex form E of the error signal e is given by

$$E = R_s - C. \tag{40}$$

By (30) and (39),

$$E = \sum_{n=0}^{\infty} (A_{2n+1}e^{j\theta_{2n+1}} + H_{2n+1}M_{2n+1}e^{-2j\alpha_{2n+1}})e^{j(2n+1)\omega t}. \tag{41}$$

Since

$$b = K_1 e + K_2 e' \tag{42}$$

the complex form B of b, in view of (26), (33), and (41) satisfies

$$A_{2n+1}e^{j(\theta_{2n+1}+\beta_{2n+1})} + H_{2n+1}M_{2n+1}e^{j(\beta_{2n+1}-2\alpha_{2n+1})} = \frac{U_{2n+1}}{G_{2n+1}}. \tag{43}$$

Equating real and imaginary parts, the amplitude A_{2n+1} and phase θ_{2n+1} of the $(2n+1)$st component of the input stabilizing signal r_s are found to satisfy

$$A_{2n+1} = \left\{ \left(\frac{U_{2n+1}}{G_{2n+1}} \right)^2 - 2\frac{U_{2n+1}}{G_{2n+1}} H_{2n+1}M_{2n+1} \cdot \cos(2\alpha_{2n+1} - \beta_{2n+1}) + (H_{2n+1}M_{2n+1})^2 \right\}^{\frac{1}{2}} \tag{44}$$

$$\sin(\theta_{2n+1} + \beta_{2n+1}) = \frac{H_{2n+1}M_{2n+1}}{A_{2n+1}}\sin(2\alpha_{2n+1} - \beta_{2n+1}). \tag{45}$$

With the constants of the system of Fig. 5 given, as well as the triangular stabilizing input to the limiter, all quantities in (44) and (45) are determined, whence these equations yield the components of the input stabilizing signal. Only elementary algebraic and trigonometric computations are involved.

VII. Graphical Determination of the Components of the Stabilizing Signal

R_{2n+1}, x_{2n+2}, and σ_{2n+1}, are introduced where

$$R_{2n+1} = \frac{A_{2n+1}}{|U_{2n+1}|} \tag{46}$$

$$x_{2n+1} = (-1)^n H_{2n+1}\left(\frac{B}{a}\right)\sin\left\{ (2n+1)\frac{\pi}{2}\left(\frac{a}{B_1}\right) \right\} \tag{47}$$

$$\sigma_{2n+1} = 2\alpha_{2n+1} - \beta_{2n+1}. \tag{48}$$

These quantities are determined by the given system constants and triangular wave. Eq. (44) now becomes

$$R_{2n+1}^2 - \left\{ x_{2n+1} - \left(\frac{\cos\sigma_{2n+1}}{G_{2n+1}} \right) \right\}^2 = \left(\frac{\sin\sigma_{2n+1}}{G_{2n+1}} \right)^2 \tag{49}$$

which represents a rectangular hyperbola on the x_{2n+1}, R_{2n+1} plane shown in Fig. 8. A construction for obtaining the hyperbola is given in Appendix II. From (46)

$$A_{2n+1} = R_{2n+1}|U_{2n+1}|. \tag{50}$$

In practice, R_{2n+1} can be obtained from the hyperbola. Since U_{2n+1} is given, (50) yields the amplitude A_{2n+1} of the $(2n+1)$st component of the input stabilizing signal r_s.

The determination of the phase angle θ_{2n+1} of the $(2n+1)$st component of r_s from (45) can be accomplished graphically as described in Appendix III, with the aid of Fig. 9.

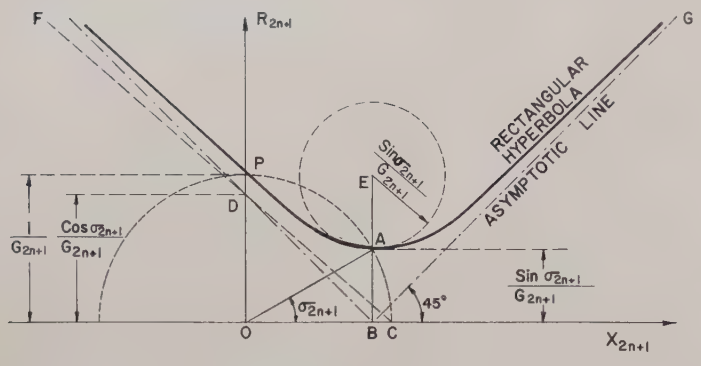


Fig. 8—Graphical solution for determination of amplitudes of harmonic components of stabilizing signal.

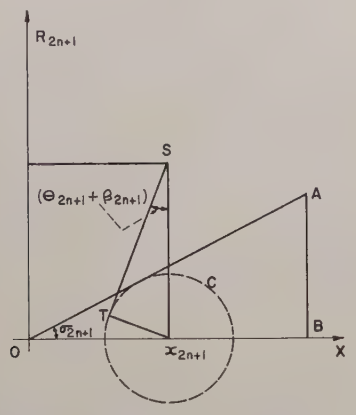


Fig. 9—Graphical solution for determination of phase angles of harmonic components of stabilizing signal.

Example

For

$$\omega = 10 \text{ rad/sec} \qquad K_1 = 1$$

$$a = 0.2 \qquad K_2 = 0.8$$

$$B = 10 \qquad \tau = 0.1,$$

Table I is now valid.

TABLE I

B_1	$r_s = A_1 \sin(\omega t+\theta_1)+A_3 \sin(3\omega t+\theta_3)$
0.2	$0.00257 \sin(\omega t-1°)+0.000718 \cos(3\omega t+19°20')$
0.3	$0.00546 \sin(\omega t-45°50')+0.00112 \cos(3\omega t+2°20')$
0.4	$0.01212 \sin(\omega t-65°40')+0.00154 \cos(3\omega t-30')$
0.6	$-0.0307 \cos(\omega t+14°20')+0.00242 \cos(3\omega t+1°20')$
1.0	$-0.0698 \cos(\omega t+10°20')+0.00398 \cos(3\omega t-40')$
2.0	$-0.1705 \cos(\omega t+8°20')+0.00795 \cos(3\omega t+40')$

Only the fundamental and third harmonic components are given in Table I since the higher harmonics are negligible. A plot of r_s for $B_1=0.3$ is shown in Fig. 10. The response c to the signal r_s is given in Table II.

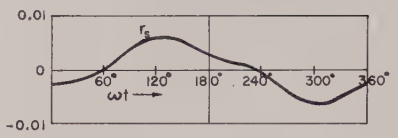


Fig. 10—Stabilizing signal r_s that yields triangular input to limiter.

TABLE II

B_1	System output c
0.2	$0.0203 \cos \omega t+0.0000632 \sin(3\omega t-143°20')$
0.3	$0.0263 \cos \omega t+0$
0.4	$0.0288 \cos \omega t-0.0000892 \sin(3\omega t-143°20')$
0.6	$0.0304 \cos \omega t-0.000190 \sin(3\omega t-143°20')$
1.0	$0.0313 \cos \omega t-0.000255 \sin(3\omega t-143°20')$
2.0	$0.0324 \cos \omega t-0.000288 \sin(3\omega t-143°\ 20')$

Conclusions

The equivalent gain 3 adequately describes the use of a stabilizing signal to remove the hunt of a given system with a nonlinearity whose output is an odd function of the input. The equivalent gain of the limiter for the sinusoidal case, working with fundamental harmonic components only, is given by a simple mathematical function 4 which can be readily used in practice. Fourier series analysis enables the designer to determine the periodic signal introduced at one point in a feedback loop to yield a given periodic stabilizing signal at another point in the loop, as shown in the theory leading to (13). The describing function is useful in determining the amplitude and frequency of a sinusoidal stabilizing signal required to remove a system hunt, as illustrated by the example of Fig. 5. For the example of Fig. 5, keeping all harmonic components, a triangular stabilizing input to the limiter yields results not much different from that of a sine wave. With the aid of the Fourier series the system input stabilizing signal r_s is given by (44) and (45) in terms of system constants and the amplitude and frequency of the triangular stabilizing input signal to the limiter. The signal r_s can be determined graphically as shown in Appendixes II and III. These techniques can be used for a wide class of systems. It is seen that signal stabilization changes the open loop gain of a closed loop system in a nonlinear manner.

Appendix I

Equivalent Gain of Limiter[6]

The input to the limiter of Fig. 2 is taken to be x, where

$$x = B_1 \sin \omega t + K_b. \qquad (51)$$

[6] This Appendix, Fig. 2 and Fig. 3 are the work of R. Sridhar, Asst. Prof. of Elec. Engrg., Purdue University.

The limiter is described by

$$y = f(x) = \begin{cases} B & x \geqq a \\ \dfrac{B}{a}x & |x| < a, \\ -B & x \leqq -a \end{cases} \tag{52}$$

where limiter input and output are x and y, respectively. We introduce angles ϕ_1 and ϕ_2 as shown in Fig. 2 so that ϕ_1, $\pi-\phi_1$, $\pi+\phi_2$ and $2\pi-\phi_2$ are associated with the corners of the limiter characteristic. The angles ϕ_1 and ϕ_2 satisfy

$$\phi_1 = \mathrm{Sin}^{-1}\frac{a-K_b}{B_1}, \qquad \phi_2 = \mathrm{Sin}^{-1}\frac{a+K_b}{B_1}. \tag{53}$$

The average value y_{av} of the output of the limiter is given by

$$\begin{aligned} y_{av} &= \frac{1}{2\pi}\Big[\int_0^{\phi_1}\frac{B}{a}(B_1\sin\theta + K_b)d\theta + \int_{\phi_1}^{\pi-\phi_1} B d\theta \\ &\quad + \int_{\pi-\phi_1}^{\pi+\phi_1}\frac{B}{a}(B_1\sin\theta + K_b)d\theta + \int_{\pi+\phi_2}^{2\pi-\phi_2} -B d\theta \\ &\quad + \int_{2\pi-\phi_2}^{2\pi}\frac{B}{a}(A\sin\theta + K_b)d\theta \\ &= \frac{1}{\pi}\left\{\frac{BB_1}{a}(\cos\phi_2 - \cos\phi_1)\right\} \\ &\quad + \frac{BK_b}{a}(\phi_1+\phi_2) + B(\phi_2-\phi_1). \end{aligned} \tag{54}$$

Now

$$\begin{aligned} &\lim_{K_b\to 0}\frac{\cos\phi_2 - \cos\phi_1}{K_b} \\ &= \lim_{K_b\to 0}\frac{\left[1-\left(\frac{a+K_b}{B_1}\right)^2\right]^{1/2} - \left[1-\left(\frac{a-K_b}{A}\right)^2\right]^{1/2}}{K_b} \\ &= -2\left(\frac{a}{B_1{}^2}\right)\left[1-\left(\frac{a}{B_1}\right)^2\right]^{-1/2}. \end{aligned} \tag{55}$$

It follows that

$$g_b = \lim_{K_b\to 0}\frac{y_{av}}{K_b} = \frac{2B}{\pi a}\mathrm{Sin}^{-1}\left(\frac{a}{B_1}\right). \tag{56}$$

Appendix II

Construction for Determining Amplitudes of Components of Stabilizing Signal

The following construction may be used to obtain the plot of the hyperbola in Fig. 8.

1) Draw a circle PAC with radius $1/G_{2n+1}$ and center at 0.
2) Draw a radius vector OA at an angle σ_{2n+1} with the x_{2n+1} axis.
3) Draw the vertical line EAB through the point A so that $\overline{EA} = \overline{AB}$.
4) Draw a circle with radius $\overline{EA}$ and center at E.
5) Draw two lines BG and BF at angles of 45° and 135° with the horizontal axis, respectively.
6) Locate D, the point of intersection of BF and the vertical axis.
7) Draw a straight line passing through C and D.
8) Draw a hyperbola with vertex at A, with radius of curvature at this point equal to $\overline{EA}$. The hyperbola should pass through the point P, the tangent at P being parallel to CD. The hyperbola thus obtained represents (48).

Appendix III

Construction for Determining Phase Angles of Components of Stabilizing Signal

θ_{2n+1} in (45) can be obtained graphically by employing the following steps (see Fig. 9).

1) Draw a circle C with center at x_{2n+1} defined by (45) such that the line OA, inclined at the angle σ_{2n+1} of (47), is tangent to the circle.
2) Let S denote the intersection of the ordinate at x_{2n+1} with the hyperbola drawn by the construction of Appendix II. Draw a tangent from S to the circle.

Then

$$\theta_{2n+1} = \underline{/TSx_{2n+1}} - \beta_{2n+1} \tag{57}$$

for the angle $\underline{/TSx_{2n+1}}$ at S.

Acknowledgment

This work was supported by the National Aeronautics and Space Administration of the United States Government.

An Analytical Approach to Root Loci*

KENNETH STEIGLITZ†

Summary—The general algebraic equations of root loci for real K are found in polar and Cartesian coordinates. A synthesis method is then suggested which leads to linear equations in the coefficients of the open-loop transfer function when closed-loop poles and their corresponding gains are specified. Equations are also found for the gain corresponding to a given point on the root locus.

A superposition theorem is presented which shows how the root loci for two open-loop functions place constraints on the locus for their product. With a knowledge of the simple lower-order loci, this theorem can be used in sketching and constructing root loci.

I. Introduction

In the usual application of the root locus technique, points are found, by a more or less trial and error procedure, at which the open-loop function is negative real. The 180° locus of the open-loop function is then sketched in the region of interest, and calibrated in terms of gain. While this graphical approach is effective in many practical problems, it is of interest to investigate the actual algebraic equations of the root loci. First of all, these equations can be used to plot, or to help sketch, the loci. Also, the equations can be used to synthesize prescribed closed-loop poles. The development will be considerably expedited by allowing the gain constant K to be both positive and negative. This idea will lead to a kind of superposition theorem for root loci, which can also be used as an aid in sketching the loci. In some cases, this approach provides exact geometrical construction procedures.

We shall be concerned with the locus of the closed-loop poles of the single-loop feedback structure shown in Fig. 1, although the results will be directly applicable

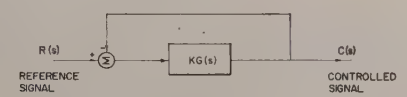


Fig. 1—The basic single-loop system.

to any system where a parameter enters linearly into the characteristic equation. The open-loop transfer function will be assumed to be a real rational function of s, such as is encountered in the analysis of linear, lumped, finite systems. We shall assume that $N(s)$, the numerator of $G(s)$, is of degree n; that the denominator $D(s)$ is of degree d; and that $N(s)$ and $D(s)$ have leading coefficients of unity. Thus, we may write

$$KG(s) = K\frac{N(s)}{D(s)} = K\frac{s^n + a_{n-1}s^{n-1} + \cdots + a_0}{s^d + b_{d-1}s^{d-1} + \cdots + b_0}$$

$$= K\frac{\sum_{k=0}^{n} a_k s^k}{\sum_{k=0}^{d} b_k s^k}, \tag{1}$$

where the a_i, b_i are real, and $a_n = b_d = 1$.

We shall now define the *root locus* corresponding to the open-loop function $G(s)$ to be the locus of the poles of the closed-loop system as the gain constant K takes on all real values, $-\infty \leq K \leq +\infty$. The closed-loop transfer function is

$$\frac{C(s)}{R(s)} = \frac{KG(s)}{1 + KG(s)}, \tag{2}$$

so that the closed-loop poles for a given value of K are given by the solutions of the equation

$$1 + KG(s) = 0, \tag{3}$$

or

$$G(s) = -\frac{1}{K}. \tag{4}$$

Since K takes on all real values, any value of s for which $G(s)$ is real will be a solution of (4). Therefore, the root locus is just the image in the s plane of the entire real axis in the G plane, and the equation of the root locus can be expressed as[1,2]

$$\text{Im}\,[G(s)] = 0, \tag{5}$$

or

$$\arg\,[G(s)] = 0°, 180°, 360°, \cdots. \tag{6}$$

We shall call those segments of the root locus for which the argument of $G(s)$ is 0°, or an even multiple of 180°, the 0° locus; and similarly, those segments for which the argument of $G(s)$ is an odd multiple of 180°, we shall call the 180° locus. Clearly, the 0° locus and the 180° locus can intersect only at infinity, or at a zero or pole of $G(s)$. Eq. (5) will give us the equation of the entire root locus, and it will remain for us to determine which segments are on the 0° locus and which are on the 180° locus.

* Received by the PGAC, September 19, 1960; revised manuscript received, February 3, 1961. This paper is based on a thesis submitted in partial fulfillment of the requirements for the M.E.E. degree at New York University, N. Y. This research was supported in part by the National Science Foundation, under whose auspices the author was a fellow.

† College of Engrg., New York University, N. Y.

[1] F. M. Reza, "Some mathematical properties of root loci for control systems design," *Trans. AIEE*, vol. 75 (*Commun. and Electronics*), pp. 103–108; March, 1956.

[2] H. Lass, "A note on the root locus method," Proc. IRE (Correspondence), vol. 44, p. 693; May, 1956.

The well-known properties of root loci are susceptible to obvious extensions under the more general definition. For instance, segments of the real axis with an odd total number of poles and zeros to the right are on the 180° locus; and the other segments are on the 0° locus. The asymptotes at infinity are at angles

$$\pm \frac{k360°}{n-d} \qquad k = 0, 1, 2, \cdots \tag{7}$$

for the 0° locus, and

$$\pm \frac{k360° + 180°}{n-d} \qquad k = 0, 1, 2, \cdots \tag{8}$$

for the 180° locus. Furthermore, these asymptotes radiate from the asymptotic center

$$\sigma_\infty = \frac{a_{n-1} - b_{d-1}}{d-n}. \tag{9}$$

We shall use the notation of Yeh,[3] and denote the root locus for an open-loop function with n zeros and d poles by $T(n, d)$.

II. The Equations of Root Loci in Polar Coordinates

We now turn to the problem of finding the general algebraic equations of root loci. First, we shall find the equations in terms of the polar coordinates shown in Fig. 2. Substituting $s = Re^{j\theta}$ into (1), we have for $G(s)$

$$G(s) = \frac{\sum_{k=0}^{n} a_k R^k e^{jk\theta}}{\sum_{l=0}^{d} b_l R^l e^{jl\theta}}. \tag{10}$$

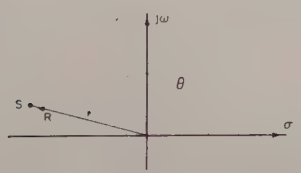


Fig. 2—The polar coordinates, R and θ.

Rationalizing (10) by multiplying by the conjugate of the denominator, we have

$$G(s) = \frac{\sum_{k=0}^{n}\sum_{l=0}^{d} a_k b_l R^{k+l} e^{j(k-l)\theta}}{\left|\sum_{l=0}^{d} b_l R^l e^{jl\theta}\right|^2}. \tag{11}$$

We may now set the imaginary part of $G(s)$ to zero as in (5) to obtain the equation of the root locus

$$\sum_{k=0}^{n}\sum_{l=0}^{d} a_k b_l R^{k+l} \sin (k-l)\theta = 0. \tag{12}$$

Since the real axis will always be part of the root locus, $R \sin \theta$ will be a factor of (12). Removing this factor, we have as the equation of the nonreal root locus

$$\sum_{k=0}^{n}\sum_{l=0}^{d} a_k b_l R^{k+l-1} \frac{\sin (k-l)\theta}{\sin \theta} = 0. \tag{13}$$

We may recognize the trigonometric functions in (13) as Tchebycheff polynomials[4] of the second kind in $\cos \theta$, defined by

$$U_{n-1}(\cos \theta) = \frac{\sin n\theta}{\sin \theta}. \tag{14}$$

Thus, the nonreal locus may be written in terms of these polynomials as

$$\sum_{k=0}^{n}\sum_{l=0}^{d} a_k b_l R^{k+l-1} U_{k-l-1}(\cos \theta) = 0. \tag{15}$$

If θ is prescribed, tables will facilitate the numerical evaluation of the $U_{k-l-1}(\cos \theta)$, and the resultant polynomial in R will have roots at the intersections of the line $\theta = \text{const.}$ with the nonreal root locus.

III. The Equations of Root Loci in Cartesian Coordinates

To find the root locus equation in terms of σ and ω, we shall use an idea that was suggested by Bendrikov and Teodorchik.[5] The idea is that of expanding $N(s)$ and $D(s)$ in a power series in $j\omega$.

$N(s)$ is analytic everywhere in the s plane, and can indeed be expanded in a Taylor series about any σ with an infinite radius of convergence to obtain the identity

$$N(\sigma + j\omega) = N(\sigma) + j\omega \frac{N'(\sigma)}{1!} + (j\omega)^2 \frac{N''(\sigma)}{2!} + \cdots + (j\omega)^n \frac{N^{(n)}(\sigma)}{n!}. \tag{16}$$

Grouping the real and imaginary components of this expression, we have

$$N(\sigma + j\omega) = \left[N(\sigma) - \omega^2 \frac{N''(\sigma)}{2!} + \cdots\right] + j\omega\left[\frac{N'(\sigma)}{1!} - \omega^2 \frac{N'''(\sigma)}{3!} + \cdots\right]. \tag{17}$$

[3] V. C. M. Yeh, "The study of transients in linear feedback systems by conformal mapping and root locus," *Trans. ASME*, vol. 76, pp. 349–361; April, 1954.

[4] See, for example, "Tables of Chebyshev Polynomials $S_n(x)$ and $C_n(x)$," Natl. Bureau of Standards Appl. Math. Ser., no. 9, Washington, D. C.; 1952.

[5] G. A. Bendrikov and K. F. Teodorchik, "The analytical theory of constructing root loci," *Avtomat. i Telemekh.* (*Automation and Remote Control*), vol. 20; March, 1959. (English Translation.)

When this is also done for $D(s)$, $G(s)$ may be written

$$G(s) = \frac{\left[N(\sigma) - \omega^2 \frac{N''(\sigma)}{2!} + \cdots\right] + j\omega\left[\frac{N'(\sigma)}{1!} - \omega^2 \frac{N'''(\sigma)}{3!} + \cdots\right]}{\left[D(\sigma) - \omega^2 \frac{D''(\sigma)}{2!} + \cdots\right] + j\omega\left[\frac{D'(\sigma)}{1!} - \omega^2 \frac{D'''(\sigma)}{3!} + \cdots\right]}. \tag{18}$$

Multiplying by the conjugate of the denominator, and setting the imaginary part equal to zero as before, we have as the equation of the nonreal root locus

$$\left[\frac{N(\sigma)D'(\sigma)}{0!1!} - \frac{N'(\sigma)D(\sigma)}{1!0!}\right] - \omega^2\left[\frac{N(\sigma)D'''(\sigma)}{0!3!} - \frac{N'(\sigma)D''(\sigma)}{1!2!} + \frac{N''(\sigma)D'(\sigma)}{2!1!} - \frac{N'''(\sigma)D(\sigma)}{3!0!}\right] + \omega^4\left[\frac{N(\sigma)}{0!}\frac{D^V(\sigma)}{5!} - \cdots\right] - \cdots = 0. \tag{19}$$

This may be written

$$Q_1(\sigma) - \omega^2 Q_3(\sigma) + \omega^4 Q_5(\sigma) - \cdots = 0, \tag{20}$$

where $Q_R(\sigma)$ is a polynomial in σ defined by

$$Q_R(\sigma) = \sum_{r=0}^{R} (-1)^r \frac{N^{(r)}(\sigma)}{r!} \frac{D^{(R-r)}(\sigma)}{(R-r)!}. \tag{21}$$

In general, $Q_R(\sigma)$ will be of degree $d+n-R$. The author has not been able to find an interpretation for all the polynomials $Q_R(\sigma)$. However, it is evident from (21) that

$$Q_{n+d}(\sigma) = (-1)^n. \tag{22}$$

It can also be shown that

$$Q_{n+d-1}(\sigma) = (-1)^n[(d-n)\sigma - (a_{n-1} - b_{d-1})], \tag{23}$$

so that if $n \neq d$,

$$Q_{n+d-1}(\sigma) = (-1)^n(d-n)(\sigma - \sigma_\infty). \tag{24}$$

Also, $Q_1(\sigma)$ can be factored by observing that the zeros of $Q_1(\sigma)$ are the roots of the equation

$$Q_1(\sigma) = N(\sigma)D'(\sigma) - N'(\sigma)D(\sigma) = 0, \tag{25}$$

or

$$N(\sigma)D(\sigma)\left[\frac{d}{d\sigma}\log\frac{N(\sigma)}{D(\sigma)}\right] = 0. \tag{26}$$

Therefore, $Q_1(\sigma)$ has zeros of appropriate orders at all the zeros of the logarithmic derivative of $G(s)$, whether they are on the root locus or not, and at the multiple poles and zeros of $G(s)$. The zeros of the logarithmic derivative represent multiple points on the phase loci of $G(s)$, and will here be called *critical points*.[6,7] By writing out the first few terms in (25), it can be seen that $Q_1(\sigma)$ has a leading coefficient of $(d-n)$, if $n \neq d$. Hence, we may write

$$Q_1(\sigma) = (d-n)(\sigma - s_{k1})(\sigma - s_{k2}) \cdots (\sigma - s_{k(n+d-1)}), \tag{27}$$

where the s_k may be complex and are solutions of (25). Ur[8] has shown that the asymptotes of the root locus for a system with no poles or zeros at infinity can be obtained by considering the asymptotes of the system $(N-D)/D$. Indeed, the root locus for $(N-D)/D$ is the same as that for N/D, since the root locus is defined by

$$\text{Im}\left\{\frac{N}{D}\right\} = \text{Im}\left\{\frac{N-D}{D}\right\} = 0. \tag{28}$$

The lower-order loci may now be written in terms of the solutions of (26) and the asymptotic center with a little more effort.

The locus for G is the same as that for $1/G$, so that $T(n, d) = T(d, n)$, and we need only consider loci for which $n \leq d$. The loci $T(1, 2)$ and $T(2, 2)$ reduce to

$$\omega^2 + (\sigma - s_{k1})(\sigma - s_{k2}) = 0, \tag{29}$$

which is in general the equation of a circle that intersects the real axis at s_{k1} and s_{k2}. The locus $T(0, 3)$ is

$$3(\sigma - s_{k1})(\sigma - s_{k2}) - \omega^2 = 0, \tag{30}$$

which is the equation of a hyperbola. The loci $T(1, 3)$ and $T(0, 4)$ are the only cubic loci, and represent the

[6] J. L. Walsh, "The Location of Critical Points of Analytic and Harmonic Functions," American Mathematical Society, New York, N. Y.; 1950.

[7] M. Marden, "The Geometry of the Zeros of a Polynomial in a Complex Variable," American Mathematical Society, New York, N. Y.; 1949.

[8] H. Ur, "Root locus properties and sensitivity relations in control systems," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-5, pp. 57–65; January, 1960.

next order of complexity after the quadratic loci. $T(1,3)$ may be written

$$\omega^2(\sigma - \sigma_\infty) + (\sigma - s_{k1})(\sigma - s_{k2})(\sigma - s_{k3}) = 0, \quad (31)$$

and $T(0, 4)$ may be written

$$\omega^2(\sigma - \sigma_\infty) - (\sigma - s_{k1})(\sigma - s_{k2})(\sigma - s_{k3}) = 0. \quad (32)$$

IV. Equations Involving the Gain Constant, K

The basic equation for the root locus, (3), has a real and imaginary part, and involves K as well as s. The equations for the root loci given above are the result of eliminating K between the two equations. If we break up $N(s)$ and $D(s)$ into real and imaginary parts as follows,

$$\begin{aligned} N(s) &= N_R(\sigma, \omega) + j\omega N_I(\sigma, \omega) \\ D(s) &= D_R(\sigma, \omega) + j\omega D_I(\sigma, \omega), \end{aligned} \quad (33)$$

(3) may be written

$$1 + K\frac{N_R(\sigma, \omega) + j\omega N_I(\sigma, \omega)}{D_R(\sigma, \omega) + j\omega D_I(\sigma, \omega)} = 0, \quad (34)$$

or

$$\begin{aligned} D_R(\sigma, \omega) + KN_R(\sigma, \omega) &= 0 \\ j\omega[D_I(\sigma, \omega) + KN_I(\sigma, \omega)] &= 0. \end{aligned} \quad (35)$$

These last two simultaneous equations fully represent the calibrated locus. In terms of polar coordinates, these equations are

$$\begin{aligned} &\sum_{k=0}^{d} b_k R^k T_k(\cos\theta) + K\sum_{k=0}^{n} a_k R^k T_k(\cos\theta) = 0 \\ &j\omega\Big[\sum_{k=1}^{d} b_k R^{k-1} U_{k-1}(\cos\theta) \\ &\qquad + K\sum_{k=1}^{n} a_k R^{k-1} U_{k-1}(\cos\theta)\Big] = 0. \end{aligned} \quad (36)$$

Here, $T_n(\cos\theta)$ is a Tchebycheff polynomial of the first kind, defined by

$$T_n(\cos\theta) = \cos n\theta. \quad (37)$$

Again, it might be mentioned that tables of these polynomials can facilitate computation. In terms of Cartesian coordinates, these equations are

$$\begin{aligned} &\Big[\frac{D(\sigma)}{0!} - \omega^2\frac{D''(\sigma)}{2!} + \cdots\Big] \\ &\quad + K\Big[\frac{N(\sigma)}{0!} - \omega^2\frac{N''(\sigma)}{2!} + \cdots\Big] = 0 \\ &j\omega\Big\{\Big[\frac{D'(\sigma)}{1!} - \omega^2\frac{D'''(\sigma)}{3!} + \cdots\Big] \\ &\quad + K\Big[\frac{D'(\sigma)}{1!} - \omega^2\frac{N'''(\sigma)}{3!} + \cdots\Big]\Big\} = 0. \end{aligned} \quad (38)$$

These equations may be used to synthesize desired closed-loop characteristics. Suppose, for example, that we require the pole-pair $s = R_1 e^{\pm j\theta_1} = \sigma_1 + j\omega_1$ to be closed-loop poles at a value of gain $K = K_1$. Then these values may be substituted into (36) or (38) to give two linear equations in the coefficients. Thus, we may specify all but two coefficients and have a closed-loop pole-pair determine these remaining two. If we specify a real closed-loop pole and its corresponding value of K, we need only substitute in the first of (36) or (38). It can be seen in fact that by specifying points in the space (s, K), we can determine any number of the coefficients by linear equations. The resultant equations will be linear in the coefficients, a_i and b_i, but will not be linear in the root positions. Thus, while the use of synthesis procedures based on (36) and (38) will lead to linear algebra in the solution for the coefficients, it will remain for the designer to bridge the gap between the unknown pole-zero positions and the coefficients in the polynomials $N(s)$ and $D(s)$.

Eqs. (35) can be solved for K as follows:

$$K = -\frac{D_R}{N_R} = -\frac{D_I}{N_I}. \quad (39)$$

Since ω is a factor of the second of (35), the second of (39) is necessarily valid only off the real axis, where $\omega \neq 0$. This restriction also applies to the second of (40), (41), and (43), which follow. We may now write K on the root locus in terms of the coordinates in the s plane and the coefficients in the open-loop transfer function. First in terms of the polar coordinates, R and θ:

$$\begin{aligned} K &= -\frac{\sum_{k=0}^{d} b_k R^k T_k(\cos\theta)}{\sum_{k=0}^{n} a_k R^k T_k(\cos\theta)} \\ &= -\frac{\sum_{k=1}^{d} b_k R^{k-1} U_{k-1}(\cos\theta)}{\sum_{k=1}^{n} a_k R^{k-1} U_{k-1}(\cos\theta)}. \end{aligned} \quad (40)$$

Or, in terms of σ and ω:

$$\begin{aligned} K &= -\frac{\dfrac{D(\sigma)}{0!} - \omega^2\dfrac{D''(\sigma)}{2!} + \omega^4\dfrac{D^{IV}(\sigma)}{4!} - \cdots}{\dfrac{N(\sigma)}{0!} - \omega^2\dfrac{N''(\sigma)}{2!} + \omega^4\dfrac{N^{IV}(\sigma)}{4!} - \cdots} \\ &= -\frac{\dfrac{D'(\sigma)}{1!} - \omega^2\dfrac{D'''(\sigma)}{3!} + \omega^4\dfrac{D^{V}(\sigma)}{5!} - \cdots}{\dfrac{N'(\sigma)}{1!} - \omega^2\dfrac{N'''(\sigma)}{3!} + \omega^4\dfrac{N^{V}(\sigma)}{5!} - \cdots}. \end{aligned} \quad (41)$$

(R, θ) and (σ, ω) in these equations are points on the root locus.

It is often of interest to find the intersections of the root locus with the $j\omega$ axis. To find these crossover points, we set $\cos\theta=0$ in (15) or $\sigma=0$ in (19). After some simplification, we have

$$(a_0b_1 - a_1b_0) - \omega^2(a_0b_3 - a_1b_2 + a_2b_1 - a_3b_0) + \omega^4(a_0b_5 - a_1b_4 + \cdots) - \cdots = 0. \quad (42)$$

The real solutions of this equation will give the values of ω at which the locus crosses the $j\omega$ axis. To find the values of K corresponding to these crossover points, we set $\cos\theta=0$ in (40) or $\sigma=0$ (41) to obtain

$$K = -\frac{b_0 - b_2\omega^2 + b_4\omega^4 - \cdots}{a_0 - a_2\omega^2 + a_4\omega^4 - \cdots} = -\frac{b_1 - b_3\omega^2 + b_5\omega^4 - \cdots}{a_1 - a_3\omega^2 + a_5\omega^4 - \cdots}, \quad (43)$$

where ω in this equation is an appropriate crossover point, a solution of (42). Note that if either $N(s)$ or $D(s)$ is a constant, both the numerator and denominator of the right-most fraction in (43) vanish, and the first expression must be used for calculations.

V. A Superposition Theorem for Root Loci

We thus have investigated the general algebraic equations of root loci. We now turn to a kind of superposition theorem for root loci; in particular, we shall show how the root loci for two open-loop functions place constraints on the locus for their product.

Theorem: Let T_1 be the root locus associated with G_1, and let T_2 be the locus associated with G_2. Then intersections of T_1 and T_2 are on the root locus associated with $G_1\cdot G_2$. Furthermore, the locus for $G_1\cdot G_2$ cannot cross the remaining parts of T_1 and T_2.

Proof: At any point which is on both T_1 and T_2, G_1 and G_2 are both real, and hence, so is $G_1\cdot G_2$. At a point on T_1 and not on T_2, G_1 is real and G_2 is not; so that $G_1\cdot G_2$ is not real and this point is not on the root locus for $G_1\cdot G_2$.

When a point on the root locus is found by this theorem, the angle of $G=G_1\cdot G_2$ at this point can be determined by adding the angles of G_1 and G_2. Thus, if a point is on the 0° locus of T_1 and the 180° locus of T_2, for instance, the point must be on the 180° locus of $G_1\cdot G_2$. On the other hand, if the point is on the 180° locus of both T_1 and T_2, it is on the 0° locus of $G_1\cdot G_2$, and so on.

This theorem is most useful when the total open-loop function can be broken up into the product of two other functions whose loci can be drawn immediately. It is important, therefore, that the user of this theorem be able to draw immediately as many loci as possible. For $T(0, 1)$, and $T(1, 1)$, or for any function which has simple poles and zeros alternating on the real axis, there is no nonreal locus. $T(0, 2)$ is a line $\sigma=$const. through the center of gravity of the poles. As pointed out by Yeh,[3] if the open-loop function consists just of an Nth order pole, the root locus coincides with the asymptotes. More generally, Lorens and Titsworth[9] state that the locus coincides with an asymptote if the pole-zero pattern is symmetric about the asymptote line extended through the asymptotic center. The Loci $T(1, 2)$ and $T(2, 2)$ are in general circles, and Fig. 3 shows these cases with enough information so each locus can be traced with a compass.

As an example of how this theorem can be used in sketching a locus, consider the open-loop function shown in Fig. 4. The asymptotes are drawn first; then the zero and poles can be divided into various groups for which simple loci can be drawn. The zero at -3 can be associated with the double pole at -5 and a circle drawn. The locus for the remaining two poles is a straight line through -3.5, and the intersections of these two loci give two points on the final locus. Moreover, this circle and line represent barriers for the final locus. Another circle-line combination is possible, and this gives two more points on the locus. When the zero is associated with other pairs of poles, it lies between them and produces no locus off the real axis, and the lines for the remaining two poles represent barriers to the locus. Thus, it is seen that the 180° locus cannot come back to meet the real axis again between the origin and the zero, and the general shape of the final locus can be sketched.

VI. Construction Procedures for Root Loci

With the device of introducing coincident pole-zero pairs, the preceding ideas give rise to construction procedures for certain loci. The simplest example of this is the hyperbolic locus $T(0, 3)$.

Consider the three-pole open-loop function shown in Fig. 5(a). Now introduce a pole and a zero which coincide, as in Fig. 5(b), so that the open-loop function and the locus is unchanged. We may now take the two real poles as one factor of $G(s)$, and the zero together with the complex pair of poles as the other. As shown in Fig. 5(c), these produce a straight line and a circle, whose intersections give a pair of points on the final locus $T(0, 3)$. By introducing another pole-zero pair along the real axis, another pair of points may be found. In this way, the locus $T(0, 3)$ can be quickly sketched as shown in Fig. 5(d). These loci may in turn be used to sketch higher-order loci.

$T(1, 3)$ can be constructed in a similar manner, as shown in Fig. 6. Here, points are located by the intersections of two circles. To construct a $T(0, 4)$ locus, such as the one shown in Fig. 7, one might introduce a real

[9] C. S. Lorens and R. C. Titsworth, "Properties of root locus asymptotes," IRE Trans. on Automatic Control, vol. AC-5, pp. 71–72; January, 1960.

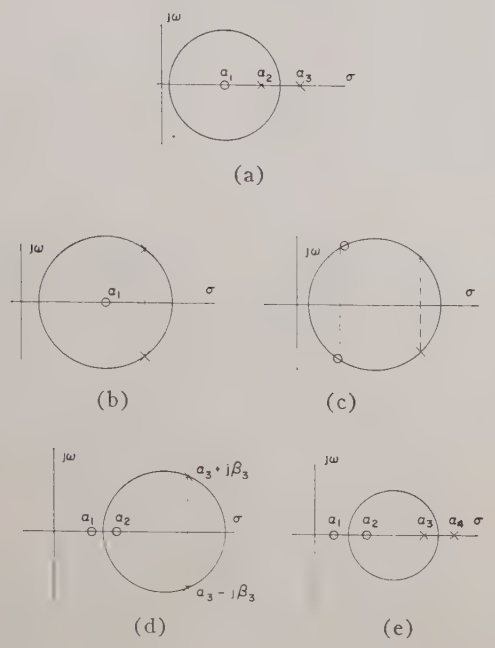


Fig. 3—The circular loci:

$(\sigma - \sigma_0)^2 + \omega^2 = R^2$

(a) $\sigma_0 = \alpha_1$, $R^2 = (\alpha_2 - \alpha_1)(\alpha_3 - \alpha_1)$

(b) $\sigma_0 = \alpha_1$

(d) $\sigma_0 = \dfrac{\alpha_3^2 + \beta_3^2 - \alpha_1\alpha_2}{2\alpha_3 - \alpha_1 - \alpha_2}$

(e) $\sigma_0 = \dfrac{\alpha_3\alpha_4 - \alpha_1\alpha_2}{\alpha_4 + \alpha_3 - \alpha_2 - \alpha_1}$, $R^2 = \sigma_0^2 + \dfrac{(\alpha_3 + \alpha_4)\alpha_1\alpha_2 - (\alpha_1 + \alpha_2)\alpha_3\alpha_4}{\alpha_4 + \alpha_3 - \alpha_2 - \alpha_1}$

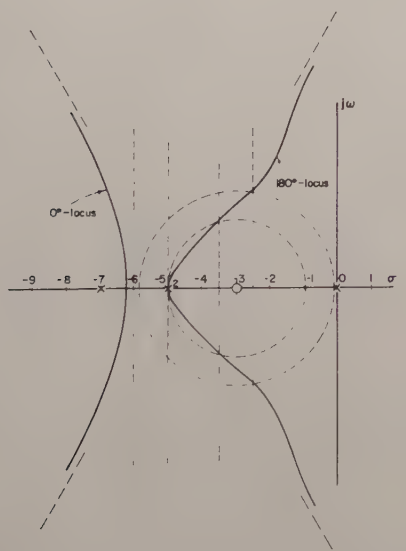


Fig. 4—The locus $T(1, 4)$ for the open-loop system

$$G(s) = \frac{(s+3)}{s(s+5)^2(s+7)}$$

sketched with the aid of the theorem of Section V.

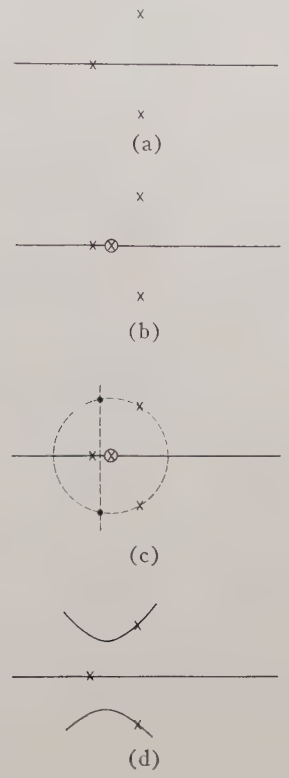


Fig. 5—A graphical procedure for constructing $T(0, 3)$. (a) The open-loop function. (b) The addition of a coincident pole and zero does not change the locus. (c) The composite loci, $T(0, 2)$ and $T(1, 2)$. (d) The final locus constructed as above.

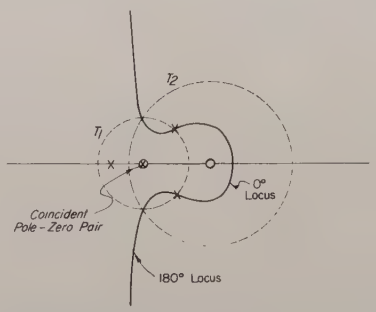


Fig. 6—Construction of the locus $T(1, 3)$ by introduction of coincident pole-zero pairs on the real axis.

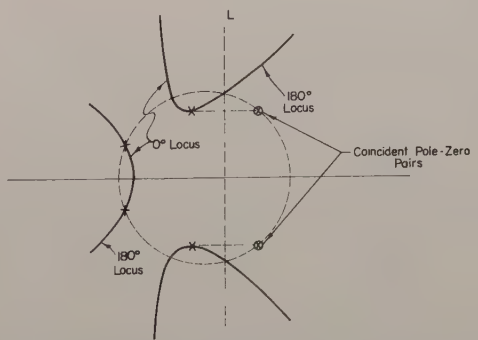


Fig. 7—Construction of the locus $T(0, 4)$ by introduction of complex pole-zero pairs.

pole-zero pair and find the intersections of hyperbolas and circles. The construction of hyperbolas can be avoided, however, in the following way. Introduce a coincident complex pair of zeros and poles with the same imaginary component as the poles in the original function. The four complex poles are now symmetrical in the line L, G is real on L, and so the line L must be part of the locus $T_1(0, 4)$. The locus $T_2(2, 2)$ is a circle, and this locates two points on the final locus.

VII. Conclusions

The general algebraic equations of root loci have been presented here in terms of the coefficients of the polynomials $N(s)$ and $D(s)$. We have seen that the specification of closed-loop poles, with their associated gains, leads to linear equations in these coefficients. The necessity of dealing with the coefficients rather than the pole-zero locations is evidently the price to be paid for linear algebra in the synthesis of closed-loop poles.

It has also been shown how the root locus for a higher-order system is restrained by the loci of its lower-order factors. A familiarity with the simple circular loci will enable the designer to use this idea as an aid in sketching loci. In certain cases, this idea leads to exact construction procedures for root loci.

Appendix

To illustrate the use of the equations for the root locus, we shall consider as an example the $T(1, 3)$ locus constructed in Fig. 6. If the zero is taken at the origin, the function is

$$G(s) = \frac{s}{[(s+1)^2+1](s+3)} = \frac{s}{s^3 + 5s^2 + 8s + 6}. \quad (44)$$

Thus,

$$\begin{aligned} a_0 &= 0 & b_0 &= 6 \\ a_1 &= 1 & b_1 &= 8 \\ & & b_2 &= 5 \\ & & b_3 &= 1. \end{aligned} \quad (45)$$

Eq. (13) then becomes

$$\sum_{l=0}^{d} b_l R^l \frac{\sin(1-l)\theta}{\sin\theta} = 6 + 5R^2 \frac{\sin(-\theta)}{\sin\theta} + R^3 \frac{\sin(-2\theta)}{\sin\theta} = 0. \quad (46)$$

or

$$2R^3 \cos\theta + 5R^2 - 6 = 0. \quad (47)$$

Since $R\cos\theta = \sigma$, this may be written simply as

$$R^2 = \frac{6}{5 + 2\sigma}, \quad (48)$$

which can be plotted quickly on polar graph paper.

Alternatively, we may find the equation of the root locus in Cartesian coordinates. Then,

$$\begin{aligned} \frac{N(\sigma)}{0!} &= \sigma & \frac{D(\sigma)}{0!} &= \sigma^3 + 5\sigma^2 + 8\sigma + 6 \\ \frac{N'(\sigma)}{1!} &= 1 & \frac{D'(\sigma)}{1!} &= 3\sigma^2 + 10\sigma + 8 \\ & & \frac{D''(\sigma)}{2!} &= 3\sigma + 5 \\ & & \frac{D'''(\sigma)}{3!} &= 1; \end{aligned} \quad (49)$$

and (19) becomes

$$[-(\sigma^3 + 5\sigma^2 + 8\sigma + 6) + \sigma(3\sigma^2 + 10\sigma + 8)] - \omega^2[\sigma(1) - 1(3\sigma + 5)] = 0, \quad (50)$$

or

$$2\sigma^3 + 5\sigma^2 - 6 + \omega^2[2\sigma + 5] = 0, \quad (51)$$

which is, of course, the same as (47). To plot this, we may wish to solve for ω^2

$$\omega^2 = -\frac{2\sigma^3 + 5\sigma^2 - 6}{2\sigma + 5}. \quad (52)$$

The gain constant on the nonreal root locus may be found by the second of (40) or (41) to be

$$K = \omega^2 - 3\sigma^2 - 10\sigma - 8. \quad (53)$$

To find the crossings of the $j\omega$ axis, for example, we have, by (48) or (52), and (53), with $\sigma = 0$

$$\omega^2 = \frac{6}{5}$$

and

$$K = -\frac{34}{5}. \quad (54)$$

Thus, these two crossover points are on the 0° locus. This may be verified by substituting directly in the system function

$$G\left(j\sqrt{\frac{6}{5}}\right) = \frac{j\sqrt{\frac{6}{5}}}{-j\frac{6}{5}\sqrt{\frac{6}{5}} - 5\frac{6}{5} + 8j\sqrt{\frac{6}{5}} + 6} = \frac{1}{34/5} = -\frac{1}{K}. \quad (55)$$

Acknowledgment

The author wishes to thank Prof. C. F. Rehberg of the Department of Electrical Engineering of New York University, under whose guidance the present research was conducted.

s-Plane Design of Compensators for Feedback Systems*

C. D. POLLAK† AND G. J. THALER‡, MEMBER, IRE

Summary—The poles and zeros of a compensator affect the *s*-plane gain and phase of the open-loop system at every point in the *s*-plane. These effects are studied for open-loop poles and zeros on the negative real axis, and a family of curves summarizes the results. A design technique is developed which permits compensation design to satisfy simultaneous specifications of root location and system gain. The method clearly defines the minimum number of compensator sections required and leads to a logical interpretation of relative needs for phase-lead and phase-lag compensators.

Statement of the Problem

COMPENSATORS are required in feedback control systems to make possible the simultaneous satisfaction of static specifications (such as the permissible threshold error due to stiction or the required system stiffness to oppose load disturbances), steady-state specifications (such as the admissible lag error in following a ramp) and dynamic performance specifications (such as rise time, bandwidth, allowable overshoot). Most compensator design procedures are based on manipulation of the open-loop transfer function. The specific manipulations usually involve trial-and-error techniques, and graphical aids are popular since they accelerate the process of converging to an acceptable compensator design. Bode-diagram[1] methods set the gain to a value which is adequate for static and steady-state accuracy specifications, then permit admustment of dynamic performance by manipulation of the magnitude and phase curves to obtain acceptable phase and gain margins. Polar-plot[1] and Nichols-Chart[1] design techniques also set the gain to a desired value and adjust the open-loop transfer-function curve to establish the resonance peak and resonant frequency of the closed-loop frequency-response curve. The root-locus method,[2–4] however, deals directly with dynamic performance by concerning itself with root locations on the *s* plane. Compensation is designed by introducing poles and zeros which force the root locus to pass through a preselected point, and the gain is adjusted to place the root at this point. Since an infinite number of pole and zero configurations can produce this result, each with a different gain value, trial-and-error is required. It has been shown[5] that compensators can be designed to place complex roots at selected points with a specified gain value and without trial-and-error, but they require that the gain be specified precisely.

The method developed in this paper solves the compensation design problem using a simplified trial-and-error method. Equations are developed which permit preliminary estimates of both gain and phase effects on the *s*-plane, and a set of curves is also provided for use in both preliminary estimates and final calculations. As a result of using these tools, the initial trial is frequently quite close to the desired answer. The method is applicable to any order system.

Theoretical Background

As a result of Chu's[6] studies in phase-angle loci and gain loci on the *s*-plane, it may be stated that any open-loop transfer function produces a phase angle and a gain value at every point on the *s*-plane. These numbers depend only on the locations of the open-loop zeros and poles, not on the gain constant of any associated amplifying equipment. To clarify this, note that certain lines on the *s*-plane are root loci, *i.e.*, loci of possible roots for the closed-loop system, and all points on such loci have a phase angle which is an odd multiple of π. Each point on the root loci has a specific gain number associated with it, and a root of the closed-loop system may be moved to a selected point on a root locus by adjusting the system-amplification constant until the necessary gain number is obtained. In computing the numerical values, it is convenient to write the loop transfer function in the form

$$G(s) = \frac{K \prod_{i}^{M} (s - z_i)}{s^n \prod_{i}^{N} (s - p_i)}, \qquad (1)$$

where K is the loop gain, z_i, p_i are the open-loop zeros and poles, $N \geqq M$, and n is the system Type Number and indicates the net number of integrations around the loop.

* Received by the PGAC, October 26, 1960; revised manuscript received, May 1, 1961.

† Lieutenant, U. S. Navy, stationed in Newport News, Va.
‡ The U. S. Naval Postgraduate School, Monterey, Calif.

[1] G. J. Thaler, "Elements of Servomechanism Theory," McGraw-Hill Book Co., Inc., New York, N. Y.; 1956.
[2] W. R. Evans, "Control System Dynamics," McGraw-Hill Book Co., Inc., New York, N. Y.; 1954.
[3] C. J. Savant, Jr., "Basic Feedback Control System Design," McGraw-Hill Book Co., Inc., New York, N. Y.; 1958.
[4] G. J. Thaler and R. G. Brown, "Analysis and Design of Feedback Control Systems," McGraw-Hill Book Co., Inc., New York, N. Y.; 1960.
[5] E. R. Ross, T. C. Warren, and G. J. Thaler, "Design of servo compensation based on the root locus method," *Trans. AIEE* (*Application and Industry*, no. 50) pp. 272–277.
[6] Y. Chu, "Synthesis of feedback control systems by phase-angle loci," *Trans. AIEE* (*Application and Industry*, no. 3), pp. 330–339; November, 1952.

From this loop transfer function, the condition for stability of the closed loop is

$$1 + \frac{K \prod_{i}^{M} (s - z_i)}{s^n \prod_{i}^{N} (s - p_i)} = 0. \tag{1a}$$

For any point s_0 which is on a root locus, the phase angle is $\underline{/G(s=s_0)} = \pi$, and the value of gain K required to locate a root at s_0 is

$$K = \frac{|s_0| \prod_{i}^{N} |s_0 - p_i|}{\prod_{i}^{M} |s_0 - z_i|}. \tag{1b}$$

For any point s_1 not on a root locus, formal substitution defines the phase angle as $\underline{/G(s=s_1)}$ and the gain number as

$$K = \frac{|s_1| \prod_{i}^{N} |s_1 - p_i|}{\prod_{i}^{M} |s_1 - z_i|}, \tag{1c}$$

To design a compensator such that a selected point s_1 becomes a root r, it is first of all necessary to introduce poles and zeros such that for the compensated transfer function G_c, the angle at point s_1 becomes

$$\underline{/G_c(s = s_1)} = (2C - 1)\pi,$$

where C is any positive integer. This merely means that a root locus must be forced to pass through s_1. It is also necessary that the gain be adjusted to move the actual root point along the root locus to s_1. An infinite number of pole-zero combinations can satisfy the angle requirement, but for each, a different gain number is obtained at s_1; and thus for each pole-zero combination, a different error coefficient[7] is obtained. Since steady-state accuracy depends on the error coefficient, only certain values of this coefficient are allowable, and it is this condition that results in trial-and-error design methods.

Assuming that the designer knows the uncompensated $G(s)$ and can select a point $s_1 = r$ at which a root is to be located, it is a simple matter to compute $\underline{/G(r)}$ using either analytic or graphical methods. Assuming also a passive compensator with poles and zeros restricted to the negative real axis, the maximum phase angle producible by a single-section compensator with one zero and one pole is obtained when the pole is at minus infinity and the zero is at the origin (or vice versa); and this maximum phase angle is

$$\beta = \mp (\pi - \tan^{-1}\omega_1/\sigma_1), \tag{2}$$

where $r = s_1 = \sigma_r + j\omega_r$. Therefore, the minimum number of compensator sections (consisting of a single zero and single pole) needed to satisfy the angle condition is

$$\textit{minimum no. of sections} = \frac{\pi - \underline{/G(r)}}{\beta}. \tag{3}$$

The gain number $K(r)$ is readily evaluated, and the error coefficient (which does not exist at $s_1 = r$ because this point is not yet on a root locus, but which may be computed formally) is found from

$$K_X = K(r) \frac{\prod_{i}^{M} (z_i)}{\prod_{i}^{N} (p_i)}, \tag{4}$$

where $K_X = K_p$, K_v, K_a, $\cdots$ depending on the system type number.

The preceding relationships are well established in the literature and are reviewed here only to provide background for the following developments.

Some Effects of Passive Compensators on the Error Coefficients

This section considers first the general concepts involved for any type system of any order, then considers in detail the second-order Type One system, developing relationships which may be used as a guide to design.

The s-plane diagram of Fig. 1 may be used to study the general case. The desired root location is r. The angle required of a lead compensator to place the root locus through r is the angle $\gamma - \alpha$. The zero and pole may be located at any positions on the negative real axis as long as this angle condition is satisfied. The gain factor of the compensator is

$$\left(\frac{z}{z_d}\right)\cdot\left(\frac{p_d}{p}\right),$$

which may be rewritten

$$\left(\frac{z}{p}\right)\left(\frac{p_d}{z_d}\right).$$

This is the factor by which the error coefficient at the root point must be modified due to the introduction of the compensator. Uncompensated, $G(s)$ is defined by (1), and the gain constant K may be defined in terms of the error coefficient K_X as

$$K \triangleq K_X \frac{\prod (p)}{\prod (z)},$$

from which

$$K_X = K \frac{\prod (z)}{\prod (p)}$$

at a selected root point r, with no compensation, the gain K evalutes to

$$|K| = \left|\frac{r^n \prod (r + p)}{\prod (r + z)}\right| = \left|\frac{r^n \prod p_d}{\prod z_d}\right|,$$

[7] J. G. Truxal, "Automatic Feedback Control System Synthesis," McGraw-Hill Book Co., Inc., New York, N. Y.; 1955.

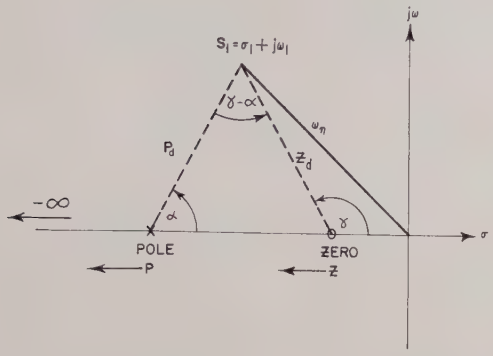


Fig. 1—*s*-plane configuration for a passive lead compensator.

and thus

$$K_X = \left(\frac{r^n \prod p_d}{\prod z_d}\right)\left(\frac{\prod z}{\prod p}\right).$$

When a single-section compensator is introduced, it has associated with it a z, p, z_d and p_d. Therefore, the error coefficient corrected for the compensator is

$$K_{XC} = K_X\left(\frac{z}{p}\right)\left(\frac{p_d}{z_d}\right).$$

When the zero is at the origin, this gain factor is zero. If the zero and pole are moved along the negative real axis toward minus infinity in such a way as to maintain the angle requirement, the ratio p_d/p approaches unity, usually without exceeding this value, and the ratio z/z_d also approaches unity, but frequently exceeds unity, *i.e.*, maximizes, depending on the value of ζ associated with the chosen root points. As an approximation which provides an upper limit for the gain value, consider the case where the compensator pole is left at minus infinity and the zero is moved along the negative real axis. The gain factor at a selected point is now defined by z/z_d, since $p_d/p = 1.0$. The angle criterion is not satisfied because the pole location is not correct, but the required finite location of the pole can only reduce the gain factor; thus the value obtained is an upper limit for the gain factor of a single section. The result is shown in Fig. 2 for ζ at the desired root location of 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8. Fig. 2 also shows the phase angle introduced by the zero. These values apply to any system and may be used as a design guide. For a lag compensator, similar results may be obtained by placing the zero at minus infinity and moving the pole. The ratio p_d/p is infinite for p at the origin, but drops to a minimum, then approaches unity as p approaches infinity.

The location of the compensator zero to produce a maximum value for the ratio z/z_d may be called z_M. It is shown in Appendix I that

$$z_M = \omega_n/\zeta, \tag{5}$$

$$\frac{z_M}{z_d} = \frac{1}{\sqrt{1-\zeta^2}}. \tag{6}$$

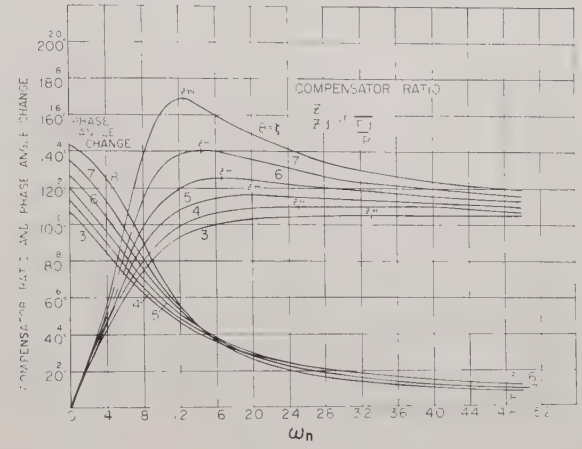


Fig. 2—Design chart for passive compensators.

Thus, the maximum possible gain increase from a single-section compensator is easily computed. This gain is never achievable because the compensator cannot be built with pole at infinity, and the value acts as a guide in that it is an upper limit.

If, in a given case where phase lead compensation is required, the pole and zero are both moved along the negative real axis toward minus infinity so as always to satisfy the angle requirement, the zero approaches a limiting position beyond which it cannot be placed if the angle criterion is to be satisfied at r. Calling this limiting location z_L, it may be noted from Fig. 1 that with a required angle $\gamma - \alpha$,

$$\tan(\gamma - \alpha) = \frac{\omega_r}{|z_L| - |\sigma_r|}, \tag{7}$$

from which

$$|z_L| = \frac{\omega_r + |\sigma_r| \tan(\gamma - \alpha)}{\tan(\gamma - \alpha)}. \tag{8}$$

Again, in practice, the zero cannot be located at z_L, and this value acts as an upper limit in guiding the design. If, for a given case $z_M < z_L$, it is possible to compensate with a single section, place the zero at or near z_M, and obtain nearly maximum gain. If $z_M > z_L$, multiple sections may be required to obtain a gain advantage.

The following simple computations permit ready estimate of the limiting gain values at a selected point r. If K_X is the uncompensated system error coefficient at r, and K_{XC} is the compensated error coefficient, then (6) is applied, and for one lead section

$$K_{XC} \leqq K_X/\sqrt{1-\zeta^2}; \tag{9}$$

for two lead sections

$$K_{XC} \leqq K_X/(1-\zeta^2); \tag{10}$$

for N lead sections

$$K_{XC} \leqq K_X/(\sqrt{1-\zeta^2})^N. \tag{11}$$

Note that these error coefficient values are upper limits. A similar development is possible for lag compensators. The curves of Fig. 2 may be used for this case also.

Design Procedure

The preceding mathematical and graphical relationships may be used to design compensation as follows: Select the location of the root r from the specifications. Evaluate $\underline{/G(r)}$, and evaluate K_X at r. Estimate the required number of sections needed to satisfy the angle criterion using (3), evaluate K_{XC} using (11) and compare K_{XC} with K_X. If $K_{XC} > K_X$, lead compensation can normally be used, otherwise one or more sections of lag compensation are required. If it is decided to use some lag compensation, it is designed first; then, the evaluation of K_{XC} is repeated preliminary to the design of the lead compensation. To design the lead compensator, evaluate z_M (using Fig. 2) and z_L, place one or more zeros as close to z_M as permitted by z_L, determine the corresponding pole locations and check K_{XC}. (K_{XC} should be at or above the specified value; sometimes a second trial is needed.) The design is thus completed, unless the additional roots introduced by the compensators are unacceptable; therefore, the locations of all roots should be checked and their effect on the dynamic response estimated.

For a second-order system, certain relationships can be established which may be used as figures of merit in selecting compensation for second-order systems, and also for higher-order systems when the selected complex roots may be considered dominant. Fig. 3 shows the root locus for a second-order Type One system, uncompensated. The basic transfer function is $G(s) = K_v/s(s\tau+1)$, and $K_v = p_{d0}p_{d1}/p_1$; but $p_{d0} = p_{d1} = \omega_n$, and $p_1 = 1/\tau_1$. Thus,

$$K_v = \omega_n^2 \tau \tag{12}$$

for the uncompensated second-order system.

When a second-order system is compensated by a lead section to produce new complex roots and the root location is defined by ω_{nc}, ζ_c, the maximum gain at the new root, as determined from Fig. 4, is

$$K_{vc} = \frac{p_{d0}p_{d1}p_{dM}}{p_1 p_M} \cdot \frac{z_M}{z_{dM}} = \frac{\omega_{nc}\tau p_{d1}}{\sqrt{1-\zeta^2}} . \tag{13}$$

p_{d1} is less than ω_{nc}, but as a first approximation let $p_{d1} \cong \omega_{nc}$; then

$$K_{vc} < \frac{\omega_{nc}^2 \tau}{\sqrt{1-\zeta^2}} . \tag{14}$$

If N sections are to be used, then

$$K_{vc} < \frac{\omega_{nc}^2 \tau}{(1-\zeta^2)^{N/2}} . \tag{15}$$

When the required value of K_{vc} is known, then the *minimum* number of lead sections is

$$N = \frac{2 \ln (\omega_{nc}^2 \tau / K_{vc})}{\ln (1-\zeta^2)} . \tag{16}$$

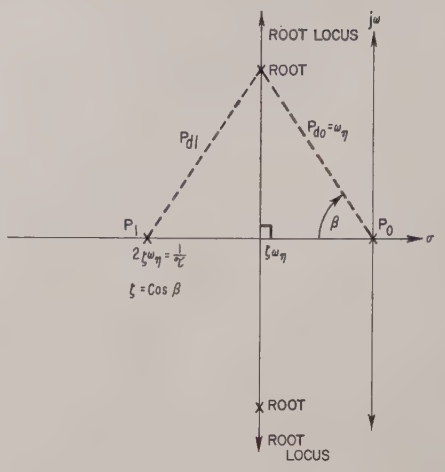


Fig. 3—Root locus for second-order, Type I system.

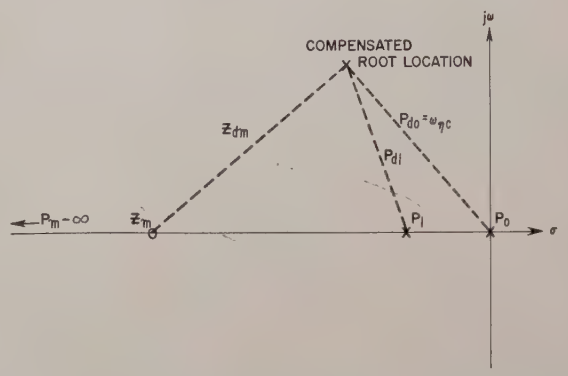


Fig. 4—K_v determination for compensated, second-order system.

In practice, the number of sections required may exceed N appreciably.

If a desired root location can be obtained with only one lead section, it is certainly possible to use more than one section. Thus, it is of interest to investigate the possibility of obtaining a gain advantage by using multiple sections. As a first step, note that for two identical sections the location of the compensator zero z_2 for maximum available gain (with poles at infinity and without regard to angle requirements) is

$$z_2 = \frac{\omega_n}{\zeta} = z_M \tag{17}$$

(see Appendix II); that is, the zero location for the double section is exactly that for the single section. If two nonidentical sections are used, the theoretical maximum gain cannot be available. This does not prove that a gain advantage necessarily accrues by use of multiple sections when one section can do the job, but it is apparent that if two zeros can be placed at z_M and the corresponding poles remain far enough out on the negative real axis, then the available gain is increased by a factor $1/\sqrt{1-\zeta^2}$ (due to the second section), which may be an appreciable advantage. In like manner, if the zeros can be kept close to z_M, a gain advantage may be available.

Illustrative Examples

Illustration 1

Given:

$$G(s) = \frac{588{,}000}{s(s+4)(s+600)}.$$

Requirements: K_v not to be reduced. Desired root location to be such that $\sigma_r = 15$, $0.5 \leqq \zeta \leqq 0.7$ [see Fig. 5(a)]. Lag networks to be avoided in this case to prevent bandwidth reduction. Attenuation ratios of lead compensators must not exceed 10.

Solution: Because the effect of the pole at $s = 600$ is negligible, the system may be considered second-order. The uncompensated system has a K_v:

$$K_v = \frac{K}{p_1 p_2} = \frac{588{,}000}{4 \cdot 600} = 245.$$

At $s_1 = \sigma_1 + j\omega_1$, for $\zeta = 0.7$,

$$K_{vu} \cong \frac{p_{d0} \cdot p_{d1}}{p_1} = \frac{15\sqrt{2}\, 18.7}{4} = 99,$$

neglecting the pole at $s = -600$. Or,

$$K_{vu} < \omega_{nc_1}{}^2 \tau = \frac{(15\sqrt{2})^2}{4} = 112.$$

With one lead section, from (14),

$$K_{vc} < \frac{\omega_{nc_1}{}^2 \tau}{\sqrt{1-\zeta^2}} < \frac{112}{\sqrt{1-0.49}} = 157.$$

With two lead sections,

$$K_{vc} < \frac{\omega_{nc_1}{}^2 \tau}{1-\zeta^2} = \frac{112}{1-0.49} = 220.$$

Since neither of these two conditions will meet requirements, examine the root $s_2 = \sigma_2 + j\omega_2$, for $\zeta = 0.5$:

$$K_{vu} < \omega_{nc_2}{}^2 \tau = \frac{(30)^2}{4} = 225.$$

With one lead section,

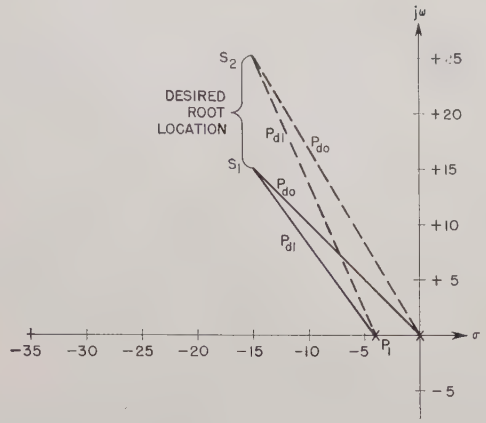
$$K_{vc} < \frac{\omega_{nc_2}{}^2 \tau}{\sqrt{1-\zeta^2}} = \frac{225}{\sqrt{1-0.25}} = 260.$$

With two lead sections,

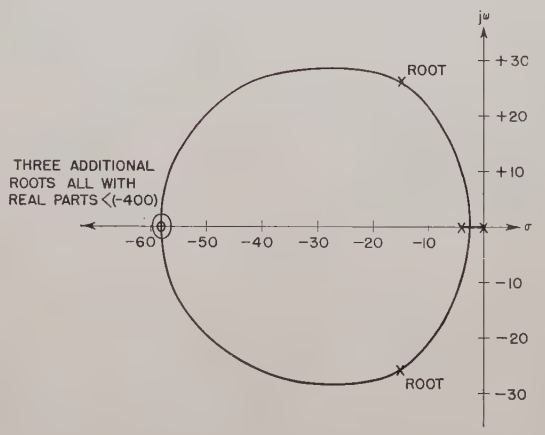
$$K_{vc} < \frac{\omega_{nc_2}{}^2 \tau}{1-\zeta^2} = \frac{225}{1-0.25} = 300.$$

The value of K_{vc} shown by one section would appear to meet specifications, but is probably too close to specifications to be realized. The value for two sections appears to be more practical. From (5),

$$z_M = \frac{\omega_n}{\zeta} = \frac{30}{0.5} = 60.$$



(a)



(b)

Fig. 5—(a) Desired root location, Illustrative Example 1. (b) Root locus after compensation, Illustrative Example 1.

Furthermore, the phase angle at $s_2 = \underline{/G(s=s_2)} = 235°$. Since two sections are to be used, each section must produce $235 - 180/2 = 27.5°$ lead. Using Fig. 2, this occurs when the z location for the compensator is at $z = 2.2$ $\omega_{nc} = 2.2(30) = 66.0$. Also from Fig. 2, a zero at $z = 2.2$ ω_{nc} produces a z/z_d ratio of 1.14. For two sections,

$$\left(\frac{z}{z_d}\right)^2 = (1.14)^2 = 1.29.$$

From this, calculate the maximum theoretical K_{vc}:

$$K_{vc} \leqq K_{vu}(1.29) = \frac{p_{d0} p_{d_1}}{p_1}(1.29) = \frac{(30)(26.7)(1.29)}{4} = 256.$$

Actual K_v after compensation will be somewhat less than this value because of the finite poles. Because the poles are finite, the zero of the compensator must be placed closer to the origin than $z = 66.0$. For a trial point, try $z = z_M = 60$. By graphical means, the pole location necessary to produce $G(s = s_2)$ is 640. This produces a phase lead attenuation of $640/60 = 10.67$, which is slightly greater than specifications permit. A second

try with the zeros at $s=59.0$ yields a pole position of 570, with a resultant attenuation ratio of $570/59=9.68$. Including the pole at $s=-600$, the final K_v is

$$K_{vc}=\frac{p_{d0}p_{d_1}p_{d_2}}{p_1p_2}\left(\frac{z_c}{z_{dc}}\right)^2\left(\frac{p_{dc}}{p_c}\right)^2$$
$$=\frac{(30)(267)(586)(59)^2(556)^2}{(4)(600)(51)^2(570)^2}=250$$

$$G_c(s)=\frac{56.0}{s(s+4)(s+600)}\cdot\frac{10^6}{}\left(\frac{s+59}{s+570}\right)^2.$$

The following points should be noted:

1) The final $K_{vc}=250$ compares to the value of 300 given by the second-order formula, (15).

2) The final $K_{vc}=250$ is within 2.4 per cent of the value given by Fig. 2. With attenuation ratios approximately equal to 10, accuracies within 5 per cent can be expected. Note also that the accuracy was achieved even though the pole at $s=-600$ was neglected until the final calculation.

3) Even though greater *compensator* ratios

$$(1/\sqrt{1-\zeta^2})^N$$

may be achieved for larger ζ locations,

$$K_{Xc}=K_{Xu}/(\sqrt{1-\zeta^2})^N$$

will generally be greater for smaller ζ locations, since the quantity K_{Xu} will dominate.

The root locus for the compensated system is shown in Fig. 5.

Illustration 2

Given:

$$G(s)=\frac{7000}{s(s+10)(s+35)}.$$

Requirements: K_v is not to be reduced. Desired root location is $r=-15\mp j15$. Use lead networks if possible, and the attenuation ratio of the lead networks must not exceed 10.

Solution:

$$K_v=7000/(10)(35)=20.$$
$$G(-15+j15)=280°.$$

Therefore, 100° of phase lead are needed.

$$K_{vu}=\frac{p_{d0}p_{d_1}p_{d_2}}{p_1p_2}=24.$$

Therefore, the gain ratio of a lead compensator must not be less than $20/24=0.834$.

From Fig. 2, for $\zeta=0.7$, one section of compensator will not work. For two identical sections, each section must produce 50° lead with a gain of $\sqrt{0.834}=0.915$. From Fig. 2, the zero may be placed at 1.25 ω_n, with gain ratio 1.39 per section. Therefore, two sections should be adequate. To keep gain at about $K_v=20$, try the zero at 0.6 ω_n. This zero produces 90° lead. To reduce the section angle to 50°, the pole must produce 40° lag. From Fig. 2, the pole goes at 1.55 ω_n. This does not work because the K_v requirement is not satisfied. Try the zero at 0.95 ω_n ($\sigma=-20$), the pole goes at 2.7 ω_n ($\sigma=-57$), the section gain ratio is 0.994 and $K_{vc}=24(0.994)^2=22.7$. Finally, with zero at $\sigma=-19$, pole at $\sigma=-47$, the gain becomes $K_v=20.02$, and the compensated transfer function is

$$G(s)=\frac{43{,}300(s+19)^2}{s(s+10)(s+35)(s+47)^2}.$$

The compensated root locus is shown in Fig. 6(a). The compensator has introduced two complex conjugate roots very near to the desired roots. The transient response is acceptable, however, as indicated in Fig. 6(b).

Illustration 3

Given: $G(s)$ as in Illustration 2.

Required: roots at $-15\mp j15$, only two lead sections permitted, with attenuation ratios less than 10. Design the lead sections from maximum K_v.

Solution: As in Illustration 2, 50° lead per section is required. From Fig. 2, the maximum zero coordinate is at 1.3 ω_n ($\sigma=-27.4$), but this violates the attenuation restriction. Try a zero at $\sigma=-26$. The pole goes at $\sigma=-220$, the attenuation ratio is $220/26=8.5$, which is acceptable and should give nearly maximum gain. The compensated transfer function is

$$G_c(s)=\frac{989{,}000(s+26)^2}{s(s+10)(s+35)(s+220)^2}$$
$$K_{vc}=39.5$$

while the maximum theoretical value is

$$K_{vc}=K_{vu}/(1-\zeta^2)=24/0.51=47.$$

A slight gain increase is possible, but not much, since the attenuation ratio per section is restricted to 10. It is interesting to note that the gain obtained is within 15 per cent of the theoretical maximum.

Illustration 4

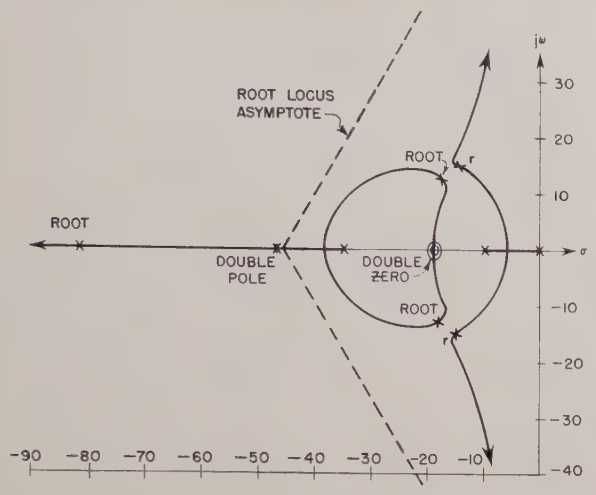
Given:

$$G(s)=\frac{150{,}000}{s(s+2)(s+10)(s+15)}.$$

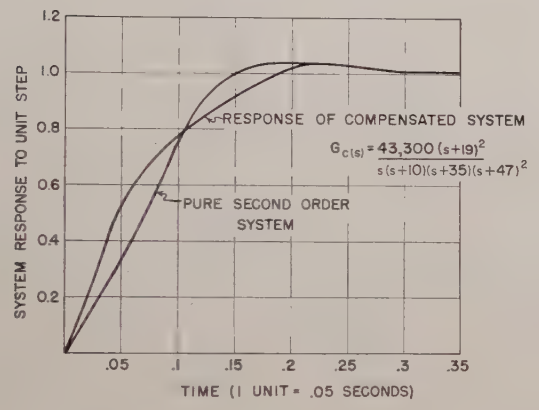
Requirements: K_v is not to be reduced. Desired root location $\sigma_r=-10$, $\zeta\geqq 0.5$. It is desired to minimize the number of compensating sections.

Solution:

$$K_v=\frac{K}{p_1p_2p_3}\frac{150{,}000}{(2)(10)(15)}=500.$$



(a)



(b)

Fig. 6—(a) Root locus after compensation, Illustrative Example 2. (b) Transient response, Illustrative Example 2.

With the root at $\sigma_r = -10$ and $\zeta = 0.5$,

$$K_{vu} = \frac{p_{d0}p_{d_1}p_{d_2}p_{d_3}}{p_1p_2p_3} = \frac{(20)(19)(17.7)(18.5)}{(2)(10)(15)} = 415.$$

The phase angle at the root is 400°. If s_1 is to be placed on the 180° phase-angle locus, 220° of phase lead compensation must be used. Because of the high K_v requirement, root locations at $\sigma_r = -10$, $\zeta > 0.5$ will not lend themselves to easier solution.

If two sections of phase lead network are used, each must produce $220/2 = 110°$ of phase lead. Fig. 2 indicates that this occurs when the zeros are placed at $s = 0.175\ \omega_n$ (for $\zeta = 0.5$). However, if lead attenuation ratios are not to exceed 10, then the poles of the lead networks must be placed at $10\ (0.175\ \omega_n) = 1.75\ \omega_n$. Fig. 2 shows that a pole at this location will produce a phase lag of 33°. The resultant net angle per section would be $110° - 33° = 77°$. Thus, it is quickly apparent that this method would be unsatisfactory.

If three sections of phase lead compensation are used, each must produce $220/3 = 73.3°$. If the zeros are placed at a point to produce 90° lead, then the poles can be located at that position to produce $(90 - 73.3) = 16.7°$ lag. From Fig. 2, the zero position, for $\zeta = 0.5$, is at $s = -0.5\omega_n = 0.5(20) = -10$; the pole position is at $s = 3.5(\omega_n) = -3.5(20) = -70$.

The lead compensator ratio

$$= \frac{z}{z_d} \cdot \frac{p_d}{p} = (0.6)\left(\frac{1}{1.15}\right) = 0.522 \text{ per section.}$$

Since three sections are to be used, this ratio must be cubed, and $(0.522)^3 = 0.142$,

$$K_{vc} = K_{vu}(0.142) = (415)(0.142) = 58.9.$$

The K_v requirement is still unsatisfied. An additional section, either lead or lag, must be added in order to increase K_{vc} to the desired value of 500. The question arises whether this may be accomplished by a fourth lead section or whether a lag section should be used. The answer may be quickly determined. If four lead sections were to be used, each must produce $220/4 = 55°$ lead. From Fig. 2, this occurs when the zero is placed at $s = -1.1\ \omega_n$, with a resultant z/z_d ratio of 1.03. Under these circumstances, K_{vc} would be given by

$$K_{vc} = K_{vu}\left(\frac{z}{z_d}\right)^4 (415)(1.03)^4 = 467,$$

and this does not account for the attenuation due to finite pole locations. Clearly, a lag compensator must be used if the compensation is to be accomplished with four sections.

From above, with three lead sections utilized to produce the proper phase angle at s_1, $K_{vc} = 58.9$. Therefore, the lag compensator must produce a gain of $500/58.9 = 8.5$. A value of 9.0 might be used to correct for any errors in calculation. To produce this gain, it is simply necessary to design a dipole with

$$\frac{z}{p} \approx 9.0.$$

Select arbitrarily a zero position of $s = -0.9$ and a pole position of $s = -0.1$. The zeros of the lead compensator sections are placed at the determined position of $s = -10$, and the poles of the lead compensators are placed in the necessary location to produce the 180° phase angle at s_1. This may be done graphically or analytically (note that this final pole position will be slightly different than the position $s = -70$ previously calculated because of the slight phase-angle effect at s_1 introduced by the lag compensator). This position is determined to be $s = -65$

$$G_c(s) = \frac{4.29\ 10^6}{s(s+2)(s+10)(s+15)} \cdot \left(\frac{s+10}{s+65}\right)^3 \cdot \left(\frac{s+0.1}{s+0.9}\right)$$

$$K_{vc} = \frac{4.29\ 10^6\ (10^3)(10^{-1})}{2\ 10\ 15\ (65^3)(9\ 10^1)} = 578.$$

Conclusions

This paper develops a method for computing the effects of compensator poles on the open-loop transfer-function gain and phase at any selected point in the s-plane. It summarizes these effects for the common case of compensator poles and zeros on the negative real axis, presenting the summary in the form of a family of curves (Fig. 2) with values of ζ for the root point used as a parameter. A design procedure has been presented which uses this information to design acceptable compensation while simultaneously considering specifications which limit the root location, gain, and number of compensation sections. The ease of application of the procedure and the accuracy of the results have been demonstrated.

Appendix I

Derivation of (5) and (6)

The location of the compensator zero to produce a maximum value for z/z_d may be called z_M. For any order system, this depends only on the chosen root location. Note that $z_d=\sqrt{\omega_r^2+(z_M-\omega_r)^2}$; then the location of z_M is determined from

$$\frac{d(z_M/z_d)}{dz_M}=\frac{d}{dz_M}\left[\omega_r^2+(z_M-z_d)^2\right]^{-1/2}=0,$$

which evaluates to

$$z_M=\omega_n/\zeta, \tag{5}$$

where ζ is the damping ratio at the chosen root location. It follows that

$$z_d=\omega_n\sqrt{1-\zeta^2}/\zeta, \tag{6}$$

and

$$\frac{z_M}{z_d}=\frac{1}{\sqrt{1-\zeta^2}}.$$

Appendix II

Derivation of (17)

Let z_2 be the location of a zero such that maximum compensator gain is obtained. Since

$$z_d=\sqrt{\omega_r^2+(z_2-\sigma_r)^2},$$

the gain per section is z_2/z_d, and for two identical sections it is $(z_2/z_d)^2$. Maximizing,

$$\frac{d(z_2/z_d)^2}{dz_2}=\frac{d}{dz_2}\left\{z_2^2/\left[\omega_r^2+(z_2-\sigma_r)^2\right]^{-1}\right\}=0. \tag{17}$$

This evaluates to $z_2=\omega_n^2/\sigma_r=\omega_n/\zeta=z_M$.

Correspondence

Correction to "Automatic Control of Three-Dimensional Vector Quantities"*

My colleague, Walter Melton, has pointed out an error in the above paper.[1] The error appears in Appendix II, pp. 55–57, and results from an improper sequence of the vector operations. The correct relation can be determined as follows:

Let the angular rate of S_d with respect to S_i be given by $\overline{W}_{id}$. Then

$$\overline{A}_{id} = D_i\overline{W}_{id}$$

where A_{id} is the angular acceleration of S_d with respect to S_i. Then, from (115)

$$\begin{aligned}\overline{A}_{id} &= D_i\overline{W}_{id}\\ &= D_d\overline{W}_{id} + \overline{W}_{id}\times\overline{W}_{id}\end{aligned}$$

so that

$$D_d\overline{W}_{id} = D_i\overline{W}_{id}.$$

This result is also demonstrated by Webster,[2] although the development is carried out in more detail. As a consequence of this result, (117)–(120), (121b), (122b) and (123b) of Part 2 may be eliminated; the corrected equations are then

$$\begin{aligned}A_1 = {} & D^2Zo - D^2Co\sin Eio\\ & - DCoDEio\cos Eio \qquad (121a)\\ A_2 = {} & D^2Co\cos Eio\sin Zo + D^2Eio\cos Zo\\ & + DCoDZo\cos Eio\cos Zo\\ & - DCoDEio\sin Eio\sin Zo\\ & - DEioDZo\sin Zo \qquad (122a)\\ A_3 = {} & D^2Co\cos Eio\cos Zo - D^2Eio\sin Zo\\ & - DCoDEio\sin Eio\cos Zo\\ & - DCoDZo\cos Eio\sin Zo\\ & - DEioDZo\cos Zo. \qquad (123a)\end{aligned}$$

That is, the factor of two appearing in the published version, the so-called "Coriolis" effect, is incorrect.

It may be noted that

$$\overline{A}^d_{id} = \begin{vmatrix}A_1\\A_2\\A_3\end{vmatrix} = \begin{vmatrix}DW_x\\DW_y\\DW_z\end{vmatrix}$$

where W_x, W_y and W_z are given in (109), (110) and (111). [Note that the sign of the last term in (111) should be negative.]

It may be recalled that the purpose of the paper[1] is to establish a vector algebra which retains the cogency of vector notation, and, in addition, provides an almost mechanical method of evaluating vectors and their components in any given set of coordinates. It is for this latter purpose that the superscript notation is introduced. However, caution is required when dealing with the problem of vector differentiation; as a result of the convention of using a column matrix to represent a vector, without including the row matrix for the unit vectors, i, k, j, there is a certain amount of ambiguity associated with the differentiation operation. That is

$$D_a\left[\,|i_aj_ak_a|\begin{vmatrix}V_x\\V_y\\V_z\end{vmatrix}\right] = |i_aj_ak_a|\begin{vmatrix}DV_x\\DV_y\\DV_z\end{vmatrix}.$$

But

$$D_b\left[\,|i_aj_ak_a|\begin{vmatrix}V_x\\V_y\\V_z\end{vmatrix}\right] = ?$$

It should not be inferred that the operation $D_b(\overline{V}^a)$ is not definable. In fact, if this operation is defined, and applied to the case above, namely $D_i(\overline{W}^d_{id})$, there results a set of steps very similar to those which evolve in Webster's work.[3] In other words, $D_a\overline{V}^a$ is defined in our algebra, but $D_b\overline{V}^a$ is not. However, we again have recourse to (115) so that

$$D_b\overline{V}^a = D_a\overline{V}^a + \overline{W}^a_{ba}\times\overline{V}^a.$$

The significance of the relationship above is illustrated in the remaining discussion, and is analogous to the problem cited in conjunction with (135) of Part 3.[4] The remaining discussion, in fact, is directed at correcting the effects of the error in Appendix II which were carried into Sections X and XI of Part 3.

In Section X, the equations through (136) are correct. Therefore, it is required to find

$$\overline{M}^e_r = J_rD_r\overline{W}^e_{ir} + \overline{W}^e_{ir}\times(J_r\overline{W}^e_{ir}).$$

Now we note that

$$D_r\overline{W}^e_{ir} = J_rD_r\overline{W}^e_{ir} + \overline{W}^e_{re}\times J_r\overline{W}^e_{ir}).$$

$$= D_e\begin{vmatrix}W_1+DGe\\W_2\\W_3+w\end{vmatrix} + \overline{W}^e_{re}\times(\overline{W}^e_{ia}+\overline{W}^e_{ae}+\overline{W}^e_{er})$$

$$= \begin{vmatrix}A_1+D_2Ge\\A_2\\A_3\end{vmatrix} + \begin{vmatrix}0\\0\\-w\end{vmatrix}\times\begin{vmatrix}W_1+DGe\\W_2\\W_3\end{vmatrix}$$

$$= \begin{vmatrix}A_1+D_2Ge+W_2w\\A_2-(W_1+DGe)w\\A_3\end{vmatrix}$$

where A_1, A_2 and A_3 may be determined as described above and we note that (139) is correct. Now (137), (138), (140), and (141) may be omitted, and using the equation for $D_r\overline{W}^e_{ir}$ developed above with (129) and (142), the correct form for (128a) is given as

$$\begin{vmatrix}M^e_{rx}\\M^e_{ry}\\M^e_{rz}\end{vmatrix} = D_i(J_r\overline{W}^e_{ir}) = \begin{vmatrix}J(A_1+D_2Ge)+(J_{sp}-J)W_2W_3+H_{sp}W_2\\JA_2+(J-J_{sp})(W_1+DGe)W_3-H_{sp}(W_1+DGe)\\J_{sp}A_3\end{vmatrix}.$$

Similarly, in Section XI, all the equations through (150) are correct, but (150a) should be written as follows:

$$\overline{A}^e_{ie} = \begin{vmatrix}A_1+D_2Ge\\A_2\\A_3\end{vmatrix}$$

where $A_1=DW_1$, $A_2=DW_2$ and $A_3=DW_3$, so that (151) should be

$$J_e\overline{A}^e_{ie} = \begin{vmatrix}J_x(A_1+D_2Ge)\\J_yA_2\\J_zA_3\end{vmatrix}$$

giving the correct form for (145a) as

$$\overline{M}^e_e = \begin{vmatrix}J_x(A_1+D_2Ge)+(J_z-J_y)W_2W_3\\J_yA_2+(J_z-J_z)W_3(W_1+DGe)\\J_zA_3+(J_y-J_x)W_2(W_1+DGe)\end{vmatrix}.$$

A. S. Lange
Bendix Systems Div.
Bendix Corp.
Ann Arbor, Mich.

* Received by the PGAC, January 9, 1961.

[1] A. S. Lange, "Automatic control of three-dimensional vector quantities—Part 2," IRE TRANS. ON AUTOMATIC CONTROL, vol. AC-5, pp. 38–57; January, 1960.

[2] A. G. Webster, "The Dynamics of Particles," Dover Publications, Inc., New York, N. Y., pp. 247–249; 1959.

[3] *Ibid.*, p. 248.

[4] A. S. Lange, "Automatic control of three-dimensional vector quantities—Part 3," IRE TRANSACTIONS ON AUTOMATIC CONTROL, vol. AC-5, pp. 106–117; June, 1960.

Additions to "Notes on the Stability Criterion for Linear Discrete Systems"*

In a preceding note [1], the author discussed a simplified form of the Schür-Cohn criterion that can be readily applied to the stability test of linear discrete systems which states the following [1]:

"A necessary and sufficient condition for the polynomial $F(z)^1 = a_0 + a_1 z + a_2 z^2 + \cdots + a_k z^k + \cdots + a_n z^n$ to have all its roots inside the unit circle is represented by the constraints $|A_k| < |B_k|$ for k odd and $|A_k| > |B_k|$ for k even ($k = 1, 2, \cdots, n$)."

The stability constants A_k and B_k are obtained from expanding the Schür-Cohn determinants and are certain combinations of the a's of the characteristic equation. It is indicated [1] that to obtain the last constraint A_n and B_n, one has to expand an nth-order determinant.

In the present note, it is shown that the last constraint $|A_n| \gtrless |B_n|$ or $|A_n| = |B_n|$ is equivalent to a certain auxiliary constraint which we will introduce. The importance of this modification lies in the fact that this auxiliary constraint is extremely simple to evaluate, while obtaining A_n and B_n involves evaluation of a determinant of order n which can be very complicated for higher-order systems.

To show this major *simplification*, the author will discuss in detail the following points.

1) Auxiliary Constraint Relation: The auxiliary constraint equation which we will introduce involves the exclusion of certain real roots outside or on the unit circule. This constraint is given by:[2]

$$F(z)\big|_{z=1} > 0 \tag{1}$$

and

$$F(z)\Big|_{z=-1} \begin{cases} > 0 & \text{for } n \text{ even} \quad (2a)^3 \\ < 0 & \text{for } n \text{ odd.} \quad (2b) \end{cases}$$

Lemma [2] *1:* If (1) and (2b) are satisfied, then there exists at least one real root of $F(z) = 0$ between plus and minus one. Also, the total number of such roots is *odd*.

Lemma 2: If (1) and (2a) are satisfied, then the total number of real roots that lie between plus and minus one is zero or *even*.

The significance of this auxiliary condition, which is very simple to test, will become evident in discussing the second point. Furthermore, (1) and (2a) or (2b) can be easily combined into one inequality using absolute values.

2) Equivalence of the Constraint $|A_n| \lessgtr |B_n|$ to the Auxiliary Constraint: To show this important point, it is simpler to distinguish between two cases.

a) n is odd: Suppose we satisfy the constraint constants up to A_{n-1} and B_{n-1}, then a generalization by Marden[4] of the Schür-Cohn criterion indicates that there exist $(n-1)$ roots inside the unit circle. The arrangement of these $(n-1)$ roots (even in number) inside the unit circle is one of two alternatives. 1) The first alternative is that, because complex roots appear in conjugate, the total number of real roots between plus and minus one is either *zero* or *even*. Now if we impose the auxiliary constraint (1) and (2b) on $F(z)$, we find that the last single real root from the constraint $|A_n| < |B_n|$ should lie inside the unit circle from Lemma 1. 2) The second alternative is when the auxiliary constraint is satisfied in addition to the first $(n-1)$ constraints, then the number of real roots between plus and minus one is either *one* or *odd*, and thus in this arrangement there exists a single complex root inside the unit circle. Since complex roots appear in conjugate, therefore the last constraint $|A_n| < |B_n|$ is necessarily satisfied. Similarly, if the auxiliary constraint is not satisfied, then this indicates a single real root outside the unit circle and thus the last constraint is also not satisfied.

For the case where $|A_n| = |B_n|$, this indicates a real root on the unit circle which is also the condition of the auxiliary constraint when (written in absolute values) equated to zero. Therefore, we have shown for *n odd* that the auxiliary constraint is equivalent to the last constraint.

b) n is even: Suppose we satisfy the constraint constants up to A_{n-1} and B_{n-1}; then this indicates that there exist $(n-1)$ roots inside the unit circle. The arrangement of these $(n-1)$ roots (odd in number) inside the unit circle is one of two alternatives. 1) The first alternative is that, because complex roots appear in conjugate, the total number of real roots between plus and minus one is either *one* or *odd*. Now if we impose the auxiliary constraint (1) and (2a) on $F(z)$, we find that the last single real root from the constraint $|A_n| > |B_n|$ should lie inside the unit circle from Lemma 2. 2) The second alternative is where the auxiliary constraint is satisfied in addition to the first $(n-1)$ constraints, then the number of real roots between plus and minus one is either *zero* or *even*, and thus in this arrangement there exists a single complex root inside the unit circle. Since complex roots appear in conjugate, therefore the last constraint $|A_n| > |B_n|$ is necessarily satisfied. Similarly, if the auxiliary constraint is not satisfied, then this indicates a single real root lies outside the unit circle and thus the last constraint is also not satisfied.

For the case when $|A_n| = |B_n|$, this indicates a real root on the unit circle which is also the condition of the auxiliary constraint (written in absolute values) when equated to zero. Therefore, we have shown for n even that the auxiliary constraint is also equivalent to the last constraint.

Therefore, for the stability test it can be concluded that the first $(n-1)$ constraints of the A's and B's should be satisfied, and the auxiliary constraint is then equivalent to the last constraint $|A_n| \lessgtr |B_n|$. This equivalence [1] has been checked for the examples discussed in this note.

3) The New Stability Criterion: Combining the previous discussions we can restate the stability criterion in a modified form as follows:

"A necessary and sufficient condition for the polynomial $F(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_k z^k + \cdots + a_n z^n$ to have all its roots inside the unit circle is represented by the constraints $|A_k| < |B_k|$ for k odd and $|A_k| > |B_k|$ for k even ($k = 1, 2, \cdots, (n-1)$) and by the following auxiliary constraint.

$$F(z)\big|_{z=1} > 0 \quad \text{and} \quad F(z)\Big|_{z=-1} \begin{cases} > 0 & n \text{ is even,} \\ < 0 & n \text{ is odd} \end{cases}$$

for $a_n > 0$.

4) Modified Schür-Cohn Criterion: From the above consideration, we can usefully modify the Schür-Cohn criterion as follows [4]:

"If for the polynomial with real coefficients

$$F(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n, a_n > 0$$

satisfying the auxiliary constraint, all the stability constants A_k and B_k ($k = 1, \cdots, n-1$) are not equal, then $F(z)$ has no zeros on the circle $|z| = 1$ and $(\mu+1)$ zeros inside the unit circle for n even and μ odd as well as for n odd and μ even. (μ is the number of variations of inequality sign in the stability constants [1, (A_1, B_1), $\cdots$, (A_{n-1}, B_{n-1})].) Furthermore, when n and μ are even and when n and μ are odd the number of zeros inside the unit circle is μ."

5) Example: We may restate the stability tests for $n = 2, 3, 4$ as follows:

a) $n = 2$, $F(z) = a_0 + a_1 z + a_2 z^2$.

Stability tests:

1) $|a_0| < |a_2|$, or $|A_1| < |B_1|$

2) $|a_0 + a_2| > |a_1|$, or $|A_2| > |B_2|$.

It should be noted that the auxiliary constraint is also (2).

b) $n = 3$, $F(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3$

1) $|a_0| < |a_3|$, $|A_1| < |B_1|$

2) $|a_0^2 - a_3^2| > |a_0 a_2 - a_1 a_3|$,[5] $|A_2| > |B_2|$.

The auxiliary constraint:

$$a_0 + a_1 + a_2 + a_3 > 0, \text{ for } a_3 > 0,$$
$$a_0 - a_1 + a_2 - a_3 < 0. \tag{3}$$

It should be noted that the two relationships of the auxiliary constraint could be combined into one by using the absolute values,

* Received by the PGAC, February 13, 1961; revised manuscript received, April 14, 1961. This research was supported by the USAF Office of Scientific Research of the ARDC under Contract No. AF 18(600)-1521.

[1] With all a_k real.

[2] It should be noted that this constraint is a necessary (not sufficient) condition for the roots of $F(z)$ to lie inside the unit circle.

[3] The constraint

$$F(z)\big|_{z=1} < 0 \text{ and } F(z)\big|_{z=-1} \begin{cases} < 0 \text{ for } n \text{ even,} \\ > 0 \text{ for } n \text{ odd,} \end{cases}$$

is also possible. However we may exclude this, without loss of generality, by always letting a_n be positive.

[4] M. Marden, "The Geometry of the Zeros of a Polynomial in a Complex Variable," American Mathematical Society, New York, N. Y., pp. 152–157; 1949.

[5] By using (1) in (2), the latter implies: $(a_0^2 - a_3^2) < (a_0 a_2 - a_1 a_3)$(2a).

i.e., $|a_0+a_2|<|a_1+a_3|$.

c) $n=4$, $F(z)=a_0+a_1z+a_2z^2+a_3z^3+a_4z^4$

1) $|a_0|<|a_4|$, $|A_1|<|B_1|$

2) $|a_0^2-a_4^2|>|a_0a_3-a_1a_4|$,[6] $|A_2|>|B_2|$

3) $|a_0^3+a_0a_2a_4+a_1a_3a_4-a_0a_4^2-a_2a_4^2-a_0a_3^2|$
$<|a_0^2a_4+a_0^2a_2+a_1^2a_4-a_0a_2a_4-a_4^3-a_0a_1a_3|=|A_3|<|B_3|$.

The auxiliary constraints:

$$a_0+a_1+a_2+a_3+a_4>0, \text{ for } a_4>0,$$
$$a_0-a_1+a_2-a_3+a_4>0 \tag{4}$$

or

$$|a_0+a_2+a_4|>|a_1+a_3|, \tag{4a}$$

for any a_4.

Summary

To illustrate the procedure for the stability test discussed in this and the former letter [1], we outline the following steps required in the stability test of the polynomial

$$F(z)=a_0+a_1z+a_2z^2+\cdots+a_nz^n=0.$$

1) Expand the determinant $|X_k+Y_k|$, for $k=1, 2, \cdots, (n-1)$ after denoting the a_i's of Y_k by b_i's. This determinant is given as follows:

$$|X_k+Y_k|=\begin{vmatrix} a_0+b_{n-k+1} & a_1+b_{n-k+2} & a_2+b_{n-k+3}\cdots a_{k-2}+b_{n-1} & a_{k-1}+b_n \\ b_{n-k+2} & a_0+b_{n-k+3} & a_1+b_{n-k+4}\cdots a_{k-3}+b_n & a_{k-2} \\ b_{n-k+3} & b_{n-k+4} & a_0+b_{n-k+4}\cdots a_{k-4} & a_{k-3} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n-2} & b_{n-1} & b_n\ 0\ 0\cdots a_1 & a_2 \\ b_{n-1} & b_n & 0\ 0\ 0\ 0 \quad a_0 & a_1 \\ b_n & 0 & 0\ 0\ 0\ 0 \quad 0 & a_0 \end{vmatrix}, \quad k=1, 2, \cdots, n-1.$$

2) After expansion, identify the stability constants A_k and B_k. This can be easily achieved by examining every term which is a product of a_i's and b_i's; if it contains an *even* number (including zero) of b_i's, then it is assigned to A_k, otherwise it is assigned to B_k.

3) After collecting the terms of A_k and B_k, replace all the b_i's by the a_i's. For the polynomial $F(z)$ to have all its roots inside the unit circle, the following constraint should be satisfied.

$$|A_k|>|B_k|, \quad k \text{ even}$$
$$|A_k|<|B_k|, \quad k \text{ odd}$$
$$\text{for } k=1, 2, \cdots, n-1$$

and the auxiliary constraint

$$F(z)\big|_{z=1}>0, \qquad F(z)\big|_{z=-1}\begin{cases}>0 & n \text{ even}\\ <0 & n \text{ odd},\end{cases}$$

for $a_n>0$

or,

$$F(1)\cdot F(-1)\begin{cases}>0, & n \text{ even}\\ <0, & n \text{ odd}\end{cases}$$

1 for any a_n.

Conclusion

In this brief note it is shown that the stability constraints for an nth-order system reduce to evaluation of $(n-1)$ determinants and two auxiliary conditions, the total number being $(n+1)$ or n if the auxiliary constraint is combined into one. By using the bilinear transformation into the ω plane, the Routh-Hurwitz constraints sometimes used for the same system are $(2n-1)$, which include n constraints on the positiveness of the constants and $(n-1)$ constraints as the determinants. However, by using the Liénard-Chipart [3] or the modified Routh-Hurwitz [3] criterion the number of these constraints could also be reduced to $(n+1)$.

E. I. Jury
Dept. of Elec. Engrg.
University of California
Berkeley, Calif.

References

[1] E. I. Jury and B. H. Bharucha, "Notes on the stability criterion for linear discrete systems," IRE Trans. on Automatic Control (Correspondence), vol. AC-6, pp. 88–90; February, 1961.
[2] D. K. Cheng, "Analysis of Linear Systems," Addison Wesley Publishing Co., Reading Mass.; 1959.
[3] F. R. Gantmacher, "Theory of Matrices," Chelsea Publishing Co., New York, N. Y., vol. 2, p. 221; 1959.
[4] A. Cohn, "Über die Anzahl der Wurzeln einer algebraischen Gleichung in einem Kreise," *Math. Z.*, vol. 14, pp. 110–148; August, 1922.
[5] E. I. Jury, "A Simplified Stability Criterion for Linear Discrete Systems," University of California, Berkeley, ERL Rept. No. 373; June, 1961.

[6] This equation can also be written as: $(a_0^2-a_4^2)^2>(a_0a_3-a_1a_4)^2$, or equivalently $[a_0A_3>a_4B_3]$. If we let a_4 be positive (as we always can by multiplying $F(z)$ by minus sign), then this relationship is written in the form $[(a_0/a_4)A_3>B_3](2a)$.

Furthermore, (1), (2) or 2a) and (3) in this case are now equivalent to the constraint: $|A_1|<|B_1|$, $B_3<0$, and $|A_3|<|B_3|$. Therefore, it is shown that for stability test of the fourth-order case, the *determinant* for obtaining (2) is *redundant*. Since we have let $a_4>0$, the auxiliary constraint for this case is then (4).

This simplification can be extended to the general case where it can be shown that the number of the determinant for obtaining the A's and the B's can be reduced. However, its derivation is deferred to a future discussion. It might be noted, however, that this additional simplification is similar to the Liénard-Chipart criterion [3] for the continuous case.

Composite Flow-Graph Technique for the Solution of Multiloop, Multisampler Sampled Systems*

Recently a number of authors have successfully used the signal flow graph technique in analyzing multiloop, multisampler, multirate sampled-data systems.[1–3]

The problem is essentially to derive the output transforms $C(z)$, $C(s)$, and $C(z, m)$, which are respectively the z transform, the Laplace transform, and the modified z transform of the output signal $c(t)$ of a sampled-data system.

The purpose of this note is to present a composite signal flow graph of a sampled-data system from which all the above mentioned transforms can be obtained by applying Mason's gain formula.[4]

The method is illustrated in the following by means of an example. Consider the multisampler system shown in Fig. 1. We shall determine the Laplace transform, the z transform, and the modified z transform of the system output $c(t)$. The signal flow graph of the system is depicted in Fig. 2, in which the samplers are eliminated and the sampler outputs are represented by artificial internal signal sources[1] X_1^* and X_5^*. By use of Mason's gain formula, the signals at nodes X_1, X_3, and X_5 are written as

$$X_1=R-G_1G_2X_1^*+G_2HX_5^* \tag{1}$$
$$X_3=G_1X_1^*-HX_5^* \tag{2}$$
$$X_5=G_1G_2X_1^*-HG_2X_5^*. \tag{3}$$

Also

$$X_2=X_1^* \tag{4}$$

and

$$X_4=X_5^*. \tag{5}$$

Taking the starred transform on both sides of (1)–(5),

$$X_1^*=R^*-(G_1G_2)^*X_1^*+(G_2H)^*X_5^* \tag{6}$$
$$X_3^*=G_1^*H_1^*-H^*X_5^* \tag{7}$$
$$X_5^*=(G_1G_2)^*X_1^*-(HG_2)^*X_5^* \tag{8}$$
$$X_2^*=X_1^* \tag{9}$$
$$X_4^*=X_5^*. \tag{10}$$

Based on (6)–(10), the sampled flow graph of the system is constructed as shown in Fig. 3. The starred or z transform of the signal at any point of the system can be determined from the sampled flow graph simply by applying Mason's formula. However, the sampled flow graph does not give the unsampled variables X_1, X_2, X_3, X_4, and X_5; these are to be obtained from the composite flow graph.

* Received by the PGAC, March 16, 1961.

[1] J. M. Salzer, "Signal flow reduction in sampled data systems," 1957 IRE WESCON Convention Record, pt. 4, pp. 166–170.
[2] R. Ash, W. H. Kim, and G. M. Kranc, "A general flow graph technique for the solution of multiloop sampled systems," *Trans. ASME*, ser. D, pp. 360–366; June, 1960.
[3] J. T. Tou, "A simplified technique for the determination of output transforms of multiloop, multisampler, variable-rate discrete-data systems," Proc. IRE, vol. 49, pp. 646–647; March, 1961.
[4] S. J. Mason, "Feedback theory—further properties of signal flow graphs," Proc. IRE, vol. 44, pp. 920–926; July, 1956.

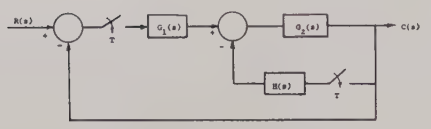


Fig. 1—Block diagram of a multisampler sampled-data system.

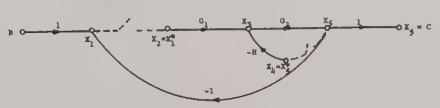


Fig. 2—Signal flow graph of sampled-data system in Fig. 1.

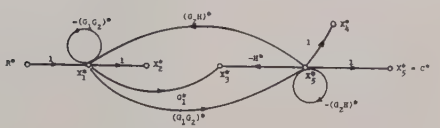


Fig. 3—Sampled flow graph for the system of Fig. 1.

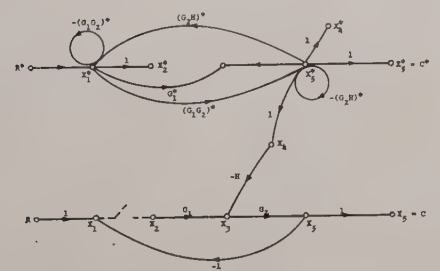


Fig. 4—Composite signal flow graph of the system of Fig. 1.

Since the sampling operation of the samplers is described by (4) and (5), this suggests that a composite signal flow graph may be constructed which is composed of the original system flow graph and the sampled flow graph with the sampling switches replaced by branches of unity gain

Once the continuous output transform is obtained, the modified z transform is determined by replacing in $C(s)$ all the starred quantities by their z-transform counterparts, and all the unstarred quantities are replaced by their counterparts in modified z transforms. Therefore,

$$C(z,m) = \frac{(G_1G_2)_m^*[1 + (G_2H)^*] - (G_1G_2)^*(G_2H)_m^*}{1 + (G_1G_2)^* + (G_2H)^*} R^* \tag{13}$$

or

$$C(z,m) = \frac{G_1G_2(z,m)[1 + G_2H(z)] - G_1G_2(z)G_2H(z,m)}{1 + G_1G_2(z) + G_2H(z)} R(z). \tag{14}$$

drawn from X_1^* to X_2 and X_5^* to X_4. The composite flow graph of the system is shown in Fig. 4, and it is seen that it has nodes that correspond to all the sampled as well as the unsampled variables.

Applying Mason's formula to this composite flow graph, the z-transform of the output signal is obtained as

$$C^* = X_5^* = \frac{(G_1G_2)^*R^*}{1 + (G_1G_2)^* + (G_2H)^*}, \tag{11}$$

and the continuous output transform is

$$C = X_5 = \frac{G_1G_2[1+(G_2H)^*] - (G_1G_2)^*G_2H}{1+(G_1G_2)^*+(G_2H)^*} R^*. \tag{12}$$

The composite flow-graph technique can be applied to systems with multiple inputs and outputs, and to systems with variable sampling rates. In the latter case, it is necessary only to replace the multirate samplers by equivalent single-rate samplers and advance and delay elements.[5]

Benjamin C. Kuo
Dept. of Elec. Engrg.
University of Illinois
Urbana, Ill.

[5] G. M. Kranc, "Input-output analysis of multirate feedback systems," IRE Trans. on Automatic Control, vol. 3, pp. 21–28; November, 1957.

Operational Analysis of Finite-Pulsed Sampled-Data Systems*

The theory of the operational analysis of the finite-pulse width system is developed in this communication. The closed-form expression of the response from such a system is described by means of several well-known operators such as the z transform, the modified z transform and the simple form of the p transform.

Finding the incremental responses and their superposition is the basic principle of the theory and it can also be applied to two-sampler systems as well as multirate sampling systems.

It is well known that sampled signals are treated in the idealized form of infinite amplitude and zero pulse width. However, signals of such forms do not exist practically, but possess finite amplitudes and nonzero pulsewidths.

Exact analysis of such finite-pulse-width systems (Fig. 1) was first attempted by G. Farmanfarma,[1–3] using the transform method. Recently E. O. Gilbert[4] has solved the same problem by means of the state-vector technique.

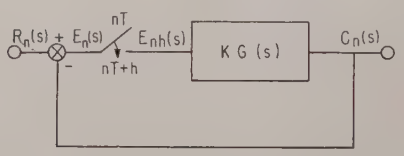


Fig. 1—Closed-loop finite-pulse width system.

In both methods, the final forms of responses at sampling instants are represented as the function of z. Paying attention to this fact, the author derived the direct-transform method to obtain the solution as a function of z without passing through the manipulation in the time domain. By this method, the solution is obtained by means of the common transformations such as the Laplace transform, z transform and modified z transform, and the final results are expressed in closed form in terms of the system constants of sampling operation and input. The introduction of other constants or new techniques of transformation is avoided as much as possible. Also, the labor for evaluating the magnitudes of transitions of variables at sampling instants, which is required in using the state-vector technique, is waived, since such evaluation is incorporated in the original derivation of equations. Another advantage of this method is that the ripples between sampling instants are given in

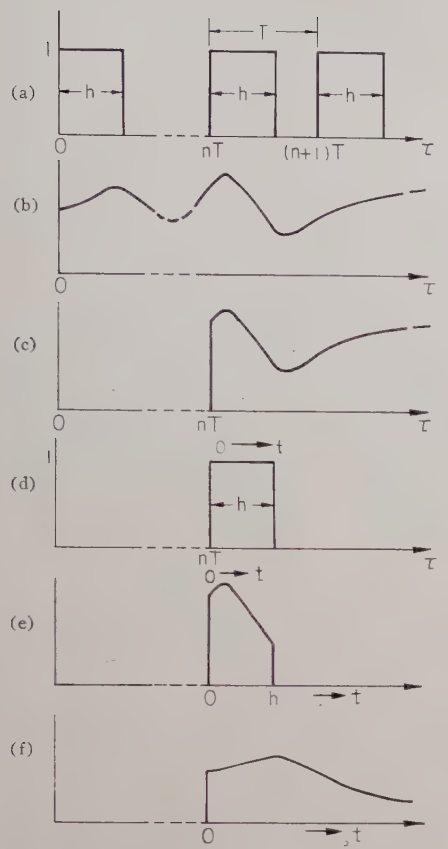
* Received by the PGAC, April 20, 1961. This research was supported by the USAF through the AF Office of Sci. Res. of the Air Res. and Dev. Command, under Contract No. AF 18(600)–1521.

[1] G. Farmanfarma, "Analysis of linear sampled data systems with finite pulse width (open loop)," *Trans. AIEE*, vol. 75 (*Commun. and Electronics*, no. 28), pp. 808–819; January, 1957.

[2] G. Farmanfarma, "General analysis and stability study of finite pulsed feedback systems," *Trans. AIEE*, vol. 77 (*Application and Industry*, no. 37), pp. 148–162; July, 1958.

[3] E. I. Jury, "Sampled-Data Control Systems," John Wiley and Sons, Inc., New York, N. Y.; 1958.

[4] E. O. Gilbert, "A method for the symbolic representation and analysis of linear periodic feedback systems," *Trans. AIEE*, vol. 79 (*Application and Industry*, no. 46), pp. 512–523; January, 1960.



(a) Sampling pattern
(b) $r(\tau)$ input
(c) $r_n'(t)$
(d) $u_h(t)$ pulse function along t axis
(e) $r_{nh}(t)$ pulsed input
(f) $c_n(t)$ output.

Fig. 2—Sampling pattern, input and output along t axis.

terms of the modified z transform, hence the derivation of equations in the time domain becomes unnecessary when the behavior of responses between sampling instants is investigated. Steady-state ripples are obtained by applying the final-value theorem to this closed-form expression of the response.[5,6] Expansion of such a closed-form expression into a power series of z^{-1} is easily performed by the digital computer. Thus the response is calculated at sampling instants as well as between sampling instants by one program on the computer, if such a closed-form expression which is applicable for both cases is derived as shown in this communication.

The theory developed in this communication is based on the superposition of the incremental response. Two kinds of z transforms are used, one of which is the ordinary z transform with impulsive sampler and the other is the z transform in power-series form that gives the proper delays to incremental responses when they are superposed.

Given the transfer function of qth order for the closed-loop finite-pulsed system (Fig. 1), the differential equation that relates the input and the output for the $(n+1)$th sampling period becomes

$$\sum_{i=0}^{q} b_i \frac{d^i c_n(t)}{dt^i} = \sum_{i=0}^{p} a_i \frac{d^i e_{nh}(t)}{dt^i} = \sum_{i=0}^{p} a_i \frac{d^i [r_{nh}(t) - c_{nh}(t)]}{dt^i} \quad 0 \le t \le T, \tag{1}$$

where the time axis t has its origin at the nth sampling instant nT as in Fig. 2. Also, the following description is adopted:

$$r_n(t) = r(t + nT), \tag{2}$$

where $r_{nh}(t)$ is the product of $r_n(t)$ and a unit pulse $u_h(t)$ and represents the pulsed input. Similar descriptions are used for $c(t)$ and $e(t)$. The Laplace transform of (1) with simple manipulation yields

$$C_n(s) = KG(s)[R_{nh}(s) - C_{nh}(s)] + \frac{1}{B(s)}\left[\sum_{i=1}^{q}\sum_{k=0}^{i-1} b_i c_n^{(k)}(0^+)s^{i-k-1}\right] - \frac{1}{B(s)}\left[\sum_{i=1}^{p}\sum_{k=0}^{i-1} a_i [r_n^{(k)}(0^+) - c_n^{(k)}(0^+)]s^{i-k-1}\right]. \tag{3}$$

$R_{nh}(s)$ is the Laplace transform of the pulsed input $r_{nh}(t)$. It may be described as the p transform[1,3] of $r_n(t)$.

$$R_{nh}(s) = L[r_{nh}(t)] = L[r_n(t)u_h(t)] = P_0^h[R_n(s)] = \frac{1}{2\pi j}\int_\Gamma R_n(p)\frac{1 - e^{-h(s-p)}}{s-p}\,dp. \tag{4}$$

$C_{nh}(s)$ is defined similarly.

While the sampler is closed, for $0 \le t \le h$, the system is continuous. Hence,

$$C_{nh}(s) = C_n(s) = C_{Hn}(s) \tag{5}$$

$$R_{nh}(s) = R_n(s), \tag{6}$$

where $C_{Hn}(s)$ represents the part of the response during the time when the sampler is closed (Fig. 2). $C_{Hn}(s)$ is obtained from (3) with the above relations, treating the system as continuous.

The response while the sampler is open can be derived, observing that $c_n(t)$ may be replaced by $c_{Hn}(t)$ when $u_h(t)$ is multiplied, since they are identical for $0 \le t \le h$ and the multiplied $u_h(t)$ makes their difference immaterial for $t > h$.

$$C_{nh}(s) = P_0^h[C_n(s)] = L[c_n(t)u_h(t)] = L[c_{Hn}(t)u_h(t)] = P_0^h[C_{Hn}(s)]. \tag{7}$$

Substituting this into (3) yields

$$C_n(s) = KG(s)P_0^h\left[\frac{R_n(s)}{1+KG(s)}\right] - \frac{\sum_{i=1}^{p}\sum_{k=0}^{i-1} a_i r_n^{(k)}(0^+)s^{i-k-1}}{B(s)} + \frac{\sum_{i=1}^{p}\sum_{k=0}^{i-1}(a_i+b_i)c_n^{(k)}(0^+)s^{i-k-1}}{B(s)} - KG(s)P_0^h\left[\frac{\sum_{i=1}^{p}\sum_{k=0}^{i-1} a_i r_n^{(k)}(0^+)s^{i-k-1}}{D(s)}\right] - KG(s)P_0^h\left[\frac{\sum_{i=1}^{q}\sum_{k=0}^{i-1}(a_i+b_i)c_n^{(k)}(0^+)s^{i-k-1}}{D(s)}\right], \tag{8}$$

where

$$D(s) = \sum_{i=0}^{q}(a_i + b_i)s^i \tag{9}$$

$$a_{p+1} = a_{p+2} = \cdots = a_q = 0. \tag{10}$$

The following relation is derived[7] which gives the response and its derivatives just before the sampling instant.

$$\sum_{i=1}^{q}\sum_{k=0}^{i-1} b_i c_n^{(k)}(0^+)s^{i-k-1} = \sum_{i=1}^{q}\sum_{k=0}^{i-1} b_i c_n^{(k)}(0^-)s^{i-k-1} + \sum_{i=1}^{p}\sum_{k=0}^{i-1} a_i [r_n^{(k)}(0^+) - c_n^{(k)}(0^+)]s^{i-k-1}. \tag{11}$$

Eq. (8) is simplified considerably by substituting the above relation into it. A new quantity $\Delta C_n(s)$, which is called the incremental response, is introduced as follows. This incremental response is the output of $KG(s)$ due to the forcing function which is applied to $KG(s)$ during the $(n+1)$th sampling period, and is given by subtracting the effect of initial conditions from $C_n(s)$.

$$\Delta C_n(s) = C_n(s) - \frac{\sum_{i=1}^{q}\sum_{k=0}^{i-1} b_i c_n^{(k)}(0^-)s^{i-k-1}}{B(s)} = KG(s)P_0^h\left[\frac{R_n(s)}{1+KG(s)}\right] + KG(s)P_0^h\left[\frac{\sum_{i=1}^{p}\sum_{k=0}^{i-1} a_i r_n^{(k)}(0^+)s^{i-k-1}}{D(s)}\right] - KG(s)P_0^h\left[\frac{\sum_{i=1}^{q}\sum_{k=0}^{i-1}(a_i+b_i)c_n^{(k)}(0^+)s^{i-k-1}}{D(s)}\right]. \tag{12}$$

This incremental response $\Delta C_n(s)$ may be rewritten as the sum of the input term and the terms due to the initial conditions of the system at $t = 0^+$. $W_{rn}(s)$ and $W_{in}(s)$ can be specified by comparing the corresponding terms in (12) and (13).

$$\Delta C_n(s) = W_{rn}(s) - \sum_{i=0}^{q-1} W_i(s)c_n^{(i)}(0^+). \tag{13}$$

The z transform of $\Delta C_n(s) = \Delta C_n^*(z)$ will give the incremental response at the sampling instants. If this incremental response is superposed for all n (from zero to infinity) with proper time lags, the sum should be equal to the total response $C^*(z)$. This operation of superposition is performed by multiplying z^{-n} to $\Delta C_n^*(z)$ in order to give the delay, and adding it from $n = 0$ to $n = \infty$. And this is identical to the operation

[5] E. I. Jury, "A note on the steady-state response of linear time-invariant systems to general periodic inputs," Proc. IRE, vol. 48, pp. 942–944; May, 1960.
[6] T. Nishimura, "Operational Analysis of Finite-Pulsed Sampled-Data Systems," Electronics Res. Lab., Univ. of Calif., Berkeley, Calif., Series No. 60, Issue No. 279, AFOSR-TN-60-510; May 10, 1960.

[7] E. I. Jury and T. Nishimura, "Analysis of finite pulsed systems with a periodically varying sampling rate and pulse width," AIEE Paper No. CP-60-866; June, 1960.

of the z transform by its definition. Thus $C^*(z)$ may be described as follows, using the operator $Zd[\ ]$ for the z transform of infinite-power-series form:

$$C^*(z) = \sum_{n=0}^{\infty} z^{-n}\Delta C_n^*(z) = Zd[\Delta C_n^*(x)] = Zd[Z[\Delta C_n(s)]]. \quad (14)$$

The response between sampling instants may be obtained by applying the modified z transform[3] to $\Delta C_n(s)$ and again in applying the z transform in infinite-power-series form. Applying the double-z transform defined in (14) to the incremental response given by (13), the following equation is derived.

$$C^*(z) = Zd[Z[\Delta C_n(s)]] = Zd[W_{rn}^*(z)] - \sum_{i=0}^{q-1} W_i^*(z)Zd[c_n^{(i)}(0^+)] = Zd[W_{rn}^*(z)] - \sum_{i=0}^{q-1} W_i^*(z)C^{(i)*}(z). \quad (15)$$

This $C^*(z)$ is given as the function of the input term and the derivatives of the response $C^{(i)*}(z)$. These derivatives are not known yet; however, before finding them the response between sampling instants will be derived by means of the modified z transform. Superposing the modified z transform of $\Delta C_n(s)$ of (13) with proper delays is performed in the same way as in (15).

$$C^*(z, m) = Zd[\Delta C_n^*(z, m)] = Zd[W_{rn}^*(z, m)] - \sum_{i=0}^{q-1} W_i^*(z, m)Zd[c_n^{(i)}(0^+)] = Zd[W_{rn}^*(z,m)] - \sum_{i=0}^{q-1} W_i^*(z, m)C^{(i)*}(z). \quad (16)$$

By the differentiation theorem[3] with respect to m, the modified z transforms of derivatives are obtained.

$$C^{(j)*}(z, m) = Zm[c^{(j)}(t)] = \frac{1}{T^j}\frac{\partial^j}{\partial m^j} C^*(z, m) \qquad j = 1, 2, \cdots, q-1. \quad (17)$$

Hence differentiating (16) j times with respect to m and dividing it by T^j yields the jth derivative

$$C^{(j)*}(z, m) = Zd[W_{rn}^{(j)*}(z, m)] - \sum_{i=0}^{q-1} W_i^{(j)*}(z, m)C^{(i)*}(z), \quad (18)$$

The z transform of the jth derivative is then

$$C^{(j)*}(z) = \lim_{m\to 0} zC^{(j)*}(z, m) = Zd[W_{rn}^{(j)*}(z)] - \sum_{i=0}^{q-1} W_i^{(j)*}(z)C^{(i)*}(z). \quad (19)$$

When all the higher derivatives for $j=1, 2, \cdots, q-1$ are derived in this way, these q equations, including (16), may be described using the matrix representation.[7]

$$C^{(j)*}(z)] = Zd(W_{rn}^{(j)*}(z))] - [W_i^{(j)*}(z)] \times C^{(i)*}(z)] \qquad i, j = 0, 1, 2, \cdots, q-1, \quad (20)$$

where $C^{(i)*}(z)]$ and $Zd(W_{rn}^{(j)*}(z))]$ are column matrices and $[W_i^{(j)*}(z)]$ is the transmission matrix which represents the terms under the summation in (19) where i refers to the column and j refers to the row of the matrix. This represents q linear equations for q unknowns, *i.e.*, $C^{(i)*}(z)$, $i=0, 1, \cdots, q-1$, and its solution yields the desired response and its derivatives at sampling instants in closed form. Substituting these $C^{(i)*}(z)$ into (16) yields the response between sampling instants, also in closed form.

The theory developed so far can naturally be applied for the open-loop system. Fig. 3 shows the response of the first-order closed-loop system when a step input is applied.

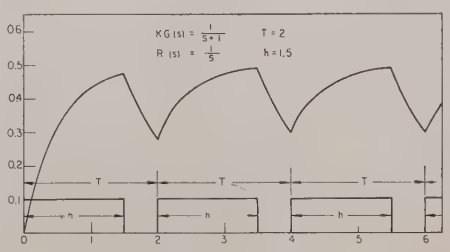


Fig. 3—Response of finite-pulse-width, first-order closed-loop system for step input.

This method is easily extended to two-sampler systems, or to multirate sampling systems.[6]

Appreciation is extended to Prof. E. I. Jury for his encouragement and suggestions.

Toshimitsu Nishimura
Dept. of Elec. Engrg.
University of California
Berkeley, Calif.

Discussion of "Optimization Based on a Square-Error Criterion with an Arbitrary Weighting Function"*

The essence of this discussion is that the above paper[1] contains a vital flaw in its basic mathematical development. Because of this flaw, the paper fails to establish that its principal results, (46), (49), (54) and (55), are valid for designing an optimum system.

It is an elementary mathematical, or logical, fact that a sufficient condition to satisfy a necessary condition is no condition at all of the original premise. Eqs. (44) and (45) of the paper constitute a sufficient, but not necessary, condition of (39), which is a valid necessary condition of an optimal system, as optimality is defined in the paper. Consequently, the purported results of the paper are not shown to be valid conditions of optimality in any sense.

Eq. (60) is a further necessary, but not sufficient, condition of a minimum or optimum. (Note: if the procedure here were extremum calculus rather than calculus of variations, the corresponding operation would result in a sufficient condition.) The validity of (60), however, does not add to the validity of (46), (49), (54) and (55) any more than the validity of (39) does, since the flaw of the derivation is in what follows (39) and (60).

To make this somewhat abstract argument plausible, the following illustration is offered. This illustration consists of a development which exactly parallels the development of the paper; even the equation numbers of the paper are used for easy reference. A valid necessary condition of optimality as given in the paper is

$$\left\{\int_{-\infty}^{+\infty} g(y)[\phi_{wrr}(\gamma - x, \tau - y) + \phi_{wrr}(\tau - x, \gamma - y)]dy - \phi_{uc_dr}(\tau, \gamma - x) - (\phi_{wc_dr}(\gamma, \tau - x)\right\}\Big|_{\tau=\gamma=0} = 0, \qquad x > 0. \quad (39)$$

Now an arbitrary function can be defined

$$\phi(\tau, \gamma)\big|_{\gamma=0,\ \text{or}\ \tau=0} = 0, \quad (A)$$

but

$$\phi(\tau, \gamma) \not\equiv 0.$$

Function ϕ has Fourier transforms with respect to both τ and γ. Combining these two,

$$\left\{\int_{-\infty}^{\infty} g(y)[\phi_{wrr}(\gamma - x, \tau - y) + \phi_{wrr}(\tau - x, \gamma - y)]dy - \phi_{wc_dr}(\tau, \gamma - x) - \phi_{wc_dr}(\gamma, \tau - x) - \phi(\tau, \gamma - x) - \phi(\gamma, \tau - x)\right\}\Big|_{\tau=\gamma=0} = 0, \qquad x > 0. \quad (39)$$

The two equations (39) are identical in view of (A). Taking the double Fourier

* Received by the PGAC, November 22, 1960.

[1] G. J. Murphy and N. T. Bold, IRE Trans. on Automatic Control, vol. AC-5, pp. 24–30; January, 1960.

transform of both sides of (39) as in the paper,

$$\left\{ \int_{-\infty}^{+\infty} e^{-j\omega'\tau} \int_{-\infty}^{+\infty} e^{-j\omega\gamma} \int_{-\infty}^{+\infty} g(y)[\phi_{wrr}(\gamma - x, \tau - y) + \phi_{wrr}(\tau - x, \gamma - y)]dy\, d\gamma d\tau \right.$$

$$\cdot \int_{-\infty}^{+\infty} e^{-j\omega\tau} \int_{-\infty}^{+\infty} e^{-j\omega\gamma}[\phi_{wc_dr}(\tau, \gamma - x) + \phi(\tau, \gamma - x)$$

$$\left. + \phi_{wc_dr}(\gamma, \tau - x) + \phi(\gamma, \tau - x)]d\gamma d\tau \right\} \Big|_{\tau=\gamma=0} = 0 \qquad x > 0, \quad (40)$$

if the integrals on the left-hand side of (40) are absolutely convergent. Changing the order of integration, which still can be done under the conditions stated in the paper, then gives

$$\left\{ \int_{-\infty}^{+\infty} g(y)[\psi_{wrr}(j\omega, j\omega')e^{-j\omega x}e^{-j\omega' y} + \psi_{wrr}(j\omega', j\omega)e^{-j\omega' x}e^{-j\omega y}]dy - e^{-j\omega x}\psi_{wc_dr}(j\omega', j\omega) - e^{-j\omega x}\psi(j\omega', j\omega) \right.$$

$$\left. - e^{-j\omega' x}\psi_{wc_dr}(j\omega, j\omega') - e^{-j\omega' x}\psi(j\omega, j\omega') \right\} \Big|_{\tau=\gamma=0} = 0 \qquad x > 0, \quad (41)$$

where ψ is the double Fourier transform of ϕ, and so on. Eq. (41) can be rewritten in the form

$$\{[G(j\omega')\psi_{wrr}(j\omega, j\omega') - \psi_{wc_dr}(j\omega, j\omega) - \psi(j\omega', j\omega)]e^{-j\omega x}$$

$$+ [G(j\omega)\psi_{wrr}(j\omega', j\omega) - \psi_{wc_dr}(j\omega, j\omega') - \psi(j\omega, j\omega')]e^{-j\omega' x}\} \Big|_{\tau=\gamma=0} = 0 \qquad x > 0. \quad (42)$$

Taking the double inverse Fourier transform of both sides of (42) gives

$$\frac{1}{4\pi^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[G(j\omega')\psi_{wrr}(j\omega, j\omega') - \psi_{wc_dr}(j\omega', j\omega) - \psi(j\omega', j\omega)]e^{-j\omega x}d\omega d\omega'$$

$$+ \frac{1}{4\pi^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}[G(j\omega)_{wrr}(j\omega', j\omega) - \psi_{wc_dr}(j\omega, j\omega') - \psi(j\omega, j\omega')]e^{-j\omega' x}d\omega d\omega' = 0 \quad x > 0, \quad (43)$$

from which the derivation of the paper concludes (note: the following (44) and (45) represent a *sufficient* condition of (43), but *not* a necessary one!) on the basis that the two integrals in (43) are identical,

$$G(j\omega')\psi_{wrr}(j\omega, j\omega') = \psi_{wc_d}(j\omega', j\omega) + \psi(j\omega', j\omega) \quad (44)$$

$$G(j\omega)\phi_{wrr}(j\omega', j\omega) = \psi_{wc_dr}(j\omega, j\omega') + \psi(j\omega, j\omega'). \quad (45)$$

The solution of these equations is

$$G_{opt}(j\omega) = \frac{\psi_{wc_dr}(j\omega, j\omega') + \psi(j\omega, j\omega')}{\psi_{wrr}(j\omega', j\omega)}, \quad (46)$$

where by (A), $\psi(j\omega, j\omega') \not\equiv 0$.

Now the above derivation follows the derivation in the paper exactly in all its steps, assumption, conclusions and logic. Nothing has been added the validity of which would be open to questions which do not apply to the derivation as it is in the paper. Yet an arbitrary function $\psi(j\omega, j\omega')$ has been introduced into the paper's result with validity equal to the validity of the result itself. In other words, an arbitrary function, subject only to the mild restrictions of (A), can be added to the paper's results without affecting the validity of these results.

This, then, amply illustrates that the results of the paper do not have any validity as conditions of an optimum system.

Furthermore, some scrutiny of (46), (49), (54), and (55) reveals that, in addition to being meaningless in substance as shown above, these equations are also meaningless in form. All of them are functions of one variable, ω, on the left-hand side, and of two variables, ω and ω', on the right-hand side. Worse, they are simple Fourier transforms on the left-hand side and double Fourier transforms on the right-hand side.

Applying the paper's results to the special case covered by Wiener's work, ω' drops out on the right-hand side of these equations, and they happen to reduce to the correct expressions. This, however, proves nothing in favor of the paper. This is only one special case, and for this case, the correct solution has been independently and validly established.

There is a possibility that other special cases, besides Wiener's, can be found, for which ω' drops out on the right-hand side of (46), (49), (54) and (55), although no examples of this have been shown so far. This merely would signify that for such special cases, these equations are not meaningless in form, but unfortunately they are still meaningless in substance. In other words, there is no basis to expect, even in such special cases, that these equations would produce optimal systems (as the paper defines optimum) unless, of course, optimality can be proven independently of the paper. Such independent proof exists in Wiener's special case.

In conclusion, then, it must be stated that the paper's principal results cannot be considered valid conditions of optimum system design. In fact, for the general case these results are meaningless even in form.

J. Zaborszky
Washington University
St. Louis, Mo.
J. W. Diesel
McDonnell Aircraft
St. Louis, Mo.

Authors' Reply[2]

The great interest of the above authors in the paper under discussion is deeply appreciated. Unfortunately, their argument is completely fallacious. Furthermore, their statements to the effect that certain of the equations in the paper are meaningless are not statements of fact but, rather, personal opinions based on their own experience and ability.

It is neither stated nor implied in the paper that satisfaction of (44) and (45) is *necessary* for minimization of the mean-weighted-square error. It is shown[3] merely that the mean-weighted-square error is made stationary by a $G(j\omega)$ that satisfies (44) and (45). Then (56) is anticipated, and (44) and (45) are accepted as sufficient (but not necessary) conditions. Later, it is shown that (62) is sufficient (but not necessary) for satisfaction of (56). Accordingly, it is concluded that a set of *sufficient* conditions for the optimum solution is

$$W(t) > 0, \quad -\infty < t < \infty;$$

$$G_{opt}(j\omega) = \frac{\psi_{wc_dr}(j\omega, j\omega')}{\psi_{wrr}(j\omega', j\omega)}.$$

The first statement in the third paragraph of the discussion by Zaborszky and Diesel is obviously absurd; (60) of the paper is no kind of condition at all—it is merely an expression for the second variation in the mean-weighted-square error. However, a sufficient condition can be obtained by equating this second variation to a constant greater than zero. Admittedly, it cannot be shown that in general in the calculus of variations if the first variation is zero and the second variation is positive, a minimum value is obtained. But for a restricted class of problems (including the problem

[2] Received by the PGAC, January 17, 1961.
[3] Murphy and Bold, *op. cit.*, p. 28.

treated in the paper under discussion), if these conditions are satisfied, a minimum is obtained. An alternative and equivalent approach to establishing the sufficiency is to show that if these conditions are satisfied, then the mean-weighted-square error assumes a larger value for all $|\beta|>0$ than it does for $\beta=0$. These two approaches to the establishing of sufficiency conditions for problems of the type under discussion are widely used.[4–8]

That the use of this alternative approach does indeed lead to the same conclusions as those presented in the paper will now be demonstrated. As in the paper, let the weighting function of the system be $g(\tau)+\beta\xi(\tau)$, and let the corresponding value of the mean-weighted-square error be $E+\Delta E$. Then ΔE is given by (34). Now if

$$G(j\omega) \equiv G_{\text{opt}}(j\omega),$$

as given in (46), the sum of the first four terms on the right-hand side of (34) is zero, with the result that

$$\Delta E = \beta^2 \int_{-\infty}^{\infty} \xi(x) \int_{-\infty}^{\infty} \xi(y)\phi_{wrr}(-x, -y)dy\,dx.$$

Since $\beta^2>0$, the sign of ΔE is the sign of

$$\int_{-\infty}^{\infty} \xi(x) \int_{-\infty}^{\infty} \xi(y)\phi_{wrr}(-x, -y)dy\,dx,$$

which is $(1/2)\partial^2\Delta E/\partial\beta^2$. It follows, therefore, that[9] if $G(j\omega)\equiv G_{\text{opt}}(j\omega)$, as given in (46), *and* $\partial^2\Delta E/\partial\beta^2|_{\beta=0}>0$, as discussed in (56)–(62), then $\Delta E>0$ for all permitted $\xi(x)$, and, hence, a minimum value of the mean-weighted-square error is obtained, as stated in the paper.

The fallacy in the argument of Zaborszky and Diesel concerning the introduction of their arbitrary function $\phi(\tau,\gamma)$ is their naive assumption that, for the class of problems treated in the paper, there *exists* a function that satisfies their definition:

$$\phi(\tau,\gamma) \not\equiv 0$$

but

$$\phi(0, \gamma) \equiv \phi(\tau, 0) \equiv 0.$$

After their "arbitrary" $\phi(\tau,\gamma)$ has been introduced, (39) of the paper becomes

$$\left\{ \int_{-\infty}^{\infty} g(y)[\phi_{wrr}(\gamma - x, \tau - y) + \phi_{wrr}(\tau - x, \gamma - y)]dy - \phi_{wc_dr}(\tau, \gamma - x) - \phi_{wc_dr}(\gamma, \tau - x) - \phi(\tau, \gamma - x) - \phi(\gamma, \tau - x) \right\}\Big|_{\tau=\gamma=0} = 0, \qquad x > 0. \quad (39A)$$

Evidently, (39A) is of the same form as (39) in the paper, except that $\phi_{wc_dr}(\tau, \gamma-x)$ and $\phi_{wc_dr}(\gamma, \tau-x)$ in the latter have been replaced with $\phi_{wc_dr}(\tau, \gamma-x)+\phi(\tau, \gamma-x)$ and $\phi_{wc_dr}(\gamma, \tau-x)+\phi(\gamma, \tau-x)$, respectively, in the former. Since there is no other difference between (39) and (39A), it follows that introducing the "arbitrary" function is equivalent to changing the desired response $c_d(t)$. That is, in effect, $c_d(t)$ has been changed to $\hat{c}_d(t)$, with $\phi_{w\hat{c}_dr}(\tau, \gamma)=\phi_{wc_dr}(\tau, \gamma)+\phi(\tau, \gamma)$. From this relation, it is evident that

$$\begin{aligned}\phi(\tau, \gamma) &= \phi_{w\hat{c}_dr}(\tau, \gamma) - \phi_{wc_dr}(\tau, \gamma)\\ &= \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} W(t)[\hat{c}_d(t+\tau) - c_d(t+\tau)]r(t+\gamma)dt\\ &= \phi_{wc_d^*r}(\tau, \gamma),\end{aligned}$$

where

$$c_d^*(t) \triangleq \hat{c}_d(t) - c_d(t).$$

Now, the requirement that

$$\phi(0, \gamma) \equiv \phi(\tau, 0) \equiv 0$$

is the requirement that

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} W(t)c_d^*(t)r(t+\gamma)dt \equiv \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} W(t)c_d^*(t+\tau)r(t)dt \equiv 0; \quad (1)$$

and the requirement that

$$\phi(\tau, \gamma) \not\equiv 0$$

is the requirement that

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} W(t)c_d^*(t+\tau)r(t+\gamma)dt \not\equiv 0. \quad (2)$$

Since (1) and (2) are not compatible, one is not able to introduce a function of the kind attempted by Zaborszky and Diesel, and hence their entire illustration "that the results of the paper do not have any validity as conditions of an optimum system" is senseless.

If (2) is replaced by

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} W(t)c_d^*(t+\tau)r(t+\gamma)dt \equiv 0, \quad (3)$$

which is compatible with (1), then the optimum solution becomes

$$G_{\text{opt}}(j\omega) = \frac{\psi_{wc_dr}(j\omega, j\omega') + \psi(j\omega, j\omega')}{\psi_{wrr}(j\omega', j\omega)},$$

where

$$\psi(j\omega, j\omega') \equiv 0.$$

This, of course, is the solution presented in (46) of the paper. Again, it should be noted that it is not even implied in the paper that such a solution always exists; the conclusion is simply that if

$$\frac{\psi_{wc_dr}(i\omega, j\omega')}{\psi_{wrr}(i\omega', j\omega)} = F(j\omega),$$

then $G_{\text{opt}}(j\omega)$ exists and is identical to $F(j\omega)$.

In that paragraph of the discussion which follows their fallacious illustration, Zaborszky and Diesel attack the form of (46), (49), (54), and (55) with the argument that one member of each of these equations is a function of only one variable, (ω), while the other member is evidently a function of two variables, $(\omega$ and $\omega')$. It seems unnecessary to point out that if

$$\psi_{wc_dr}(j\omega, j\omega') = A(j\omega)B(j\omega')$$

and

$$\psi_{wrr}(j\omega', j\omega) = C(j\omega)B(j\omega'),$$

then

$$G_{\text{opt}}(j\omega) = \frac{\psi_{wc_dr}(j\omega, j\omega')}{\psi_{wrr}(j\omega', j\omega)} = \frac{A(j\omega)}{C(j\omega)}$$

is a function of ω only. Furthermore, as is indicated in the paper, an optimum solution exists *if* (but not *only* if) the integrands in (43) are identically zero, and only if these integrands are identically zero is the optimum solution given by (46). The condition that the integrands be zero is, of course, equivalent to the condition that the ratio of $\psi_{wc_dr}(j\omega, j\omega')$ to $\psi_{wrr}(j\omega', j\omega)$ be independent of ω', as in the above example.

Apparently, it is also believed by Zaborszky and Diesel that Fourier transforms possess some magical properties or characteristics that distinguish them from ordinary functions of complex variables, which is, of course, not true. No other explanation for their statement "Worse, they are simple Fourier transforms on the left-hand side and double Fourier transforms on the right-hand side" is apparent.

That the results obtained in the paper agree with the well-known solution in the special case where the weighting function $W(t)$ is a constant has already been shown[10] by the author. Contrary to the attitude of Zaborszky and Diesel, the fact that the correct solution is obtained when the method presented in the paper is applied to the special case covered by Wiener's work is due to more than a happy coincidence. It serves, in fact, to demonstrate that for one subclass of problems in the class to which the results of the paper are applicable, the method presented in the paper does indeed yield a correct solution.

In conclusion, it must be stated that the results presented in the paper are valid and can be used to obtain the optimum system in all cases in which the stated conditions are fulfilled. Eqs. (46), (49), (54), and (55) are meaningless *neither* in form *nor* in substance; and the contentions of Zaborszky and Diesel that their "derivation follows the derivation in the paper exactly in all its steps, assumptions, conclusions, and logic" and that "Nothing has been added the validity of which would be open to questions which do not apply to the derivation as it is in the paper" are false.

G. J. Murphy
Elec. Engrg. Dept.
Northwestern University
Evanston, Ill.

[4] J. G. Truxal, "Automatic Feedback Control System Synthesis," McGraw-Hill Book Co., Inc., New York, N. Y., p. 478; 1955.

[5] G. C. Newton, Jr., L. A. Gould, and J. F. Kaiser, "Analytical Design of Linear Feedback Controls," John Wiley and Sons, Inc., New York, N. Y., p. 145; 1957.

[6] J. H. Laning, Jr., and R. H. Battin, "Random Processes in Automatic Control," McGraw-Hill Book Co., Inc., New York, N. Y., pp. 300–301; 1956.

[7] Y. W. Lee, "Statistical Theory of Communication," John Wiley and Sons, Inc., New York, N. Y., pp. 367–369; 1960.

[8] R. S. Burington and C. C. Torrance, "Higher Mathematics," McGraw-Hill Book Co., Inc., New York, N. Y., p. 788; 1939.

[9] Here, as in the paper under discussion, only those cases in which the various operations required can be justified are considered.

[10] J. B. Cruz, Jr., and G. J. Murphy, "Comments on 'Optimization based on a square-error criterion'," IRE Trans. on Automatic Control (Correspondence), vol. AC-5, pp. 328–329; September, 1960.

Russian Contributions to Control Theory*

In recent years the field of optimal control has attracted a great deal of attention, both in this country and in the Soviet Union. Unfortunately, a good deal of the Russian work has not been available in English. Since two fundamental Russian papers have now been translated, I believe it would be of interest to many readers to know that these translations are available.

The first of the papers, by R. V. Gamkrelidze, is entitled "The Theory of Time-Optimal Processes in Linear Systems." It originally appeared in the *Izvestia Akademii Nauk SSSR, Ser. Mat.*, vol. 22 (1958), pp. 449–474. It is available in English as a report of the Department of Engineering of the University of California, Los Angeles. The report number is 61-7, dated January, 1961, and is available upon request.

The second paper, by V. G. Boltyanskii, R. V. Gamkrelidze, and L. S. Pontryagin, is entitled "The Theory of Optimal Processes. I. The Maximum Principle." It appeared in the same journal as the above paper, vol. 24, no. 1 (1960) pp. 3–42. It will be available this summer in volume 17 of the *American Mathematical Society Translations*. This volume contains a number of other translations, and is sold by the American Mathematical Society, 190 Hope Street, Providence 6, R. I.

The second of these papers is the definitive work on the maximum principle by Pontryagin and his students, and summarizes all their results (for the nonlinear case) to date. The first paper, as its title implies, deals only with linear systems, and forms excellent background material for the other paper.

Lucien W. Neustadt
Space Technology Labs., Inc.
Los Angeles, Calif.

* Received by the PGAG, April 10, 1961.

Comments on "Mathematical Aspects of the Synthesis of Linear Minimum Response-Time Controllers"*

I should like to comment on Lee's paper,[1] and, in particular, on the discussion by Hopkin. As is pointed out in the discussion, (15)—$\partial f/\partial t_i = 0$ for $i = 1, \cdots, k$—is satisfied at extremal points in the *interior* of the region R, given by $0 \leq t_1 \leq t_2 \leq \cdots \leq t_k \leq t_r$. If $F(t_1, \cdots, t_k)$ takes on its minimum on the boundary of R, this condition need not be satisfied, and the remainder of the argument breaks down. The importance of this lies in the fact that the latter situation (a minimum on the boundary) is not just an isolated possibility, but is rather the one which is most likely to occur.

Suppose the matrix A has complex, or multiple real roots, or is time varying. Then the minimum number of switching times is not known beforehand. Since k must not be less than this number, a reasonable guess for k would be a large integer. But if k is greater than the optimal number of switching times, the minimum of $F(t_1, \cdots, t_k)$ will occur when a certain number of the t_i are equal (the equalities expressing this redundancy will in general not be unique—so that the minimum occurs at a number of boundary points). But if $t_i = t_{i+1}$, the point $(t_1, \cdots, t_i, t_{i+1}, \cdots, t_k)$ lies on the boundary of R. Furthermore, there may be a local minimum in the interior of R (at which $\partial f/\partial t_i = 0$), which may mask the absolute minimum.

The author also makes the statement that his Theorem I can be proved by a simple extension of arguments by Bellman, *et al.*, [1]. A general proof of this theorem—which is by no means trivial—is found in LaSalle [2]. However, if the author restricts himself to *normal* systems (see [2])—and his discussion of general linear processes implicitly assumes this—the argument is indeed simple.

A different synthesis technique for the linear time optimal control problem, based on a method of successive approximations, has been developed, and has been published in [3].

Lucien W. Neustadt
Space Technology Labs., Inc.
Los Angeles, Calif.

References

[1] R. Bellman, I. Glicksberg, and O. Gross, "On the bang-bang control problem," *Quart. Appl. Math.*, vol. 14, pp. 11–18; April, 1956.
[2] J. P. LaSalle, "The Bang-Bang Principle," presented at IFAC, Moscow, USSR; July, 1960. Also available as RIAS Tech. Rept. 59-5, RIAS, Baltimore, Md.
[3] L. W. Neustadt, "Synthesizing time-optimal control systems," *J. Math. Analysis and Applications*, vol. 1, no. 3–4, pp. 484–493; December, 1960.

Author's Comment[2]

I am pleased to note the interest expressed here and elsewhere in finding a procedure for synthesizing a time-optimal control. Many ideas have been indicated as to how to solve this problem. The procedure which I have indicated involves an iteration procedure which must converge to a minimum in a space of switching times of unknown dimension and which resembles a high-dimensional wedge. The iterative procedure not only must search for a minimum, but also must find the absolute minimum. Depending on the dimension of the problem, this may not be satisfactory because of the amount of computation required. I therefore wait with a great deal of enthusiasm for any procedure which will circumvent or partly alleviate some of the computation.

Smith[3] has demonstrated that for a system with stable distinct real characteristic roots, the procedure using switching times can be used to synthesize a time-optimal control.

E. B. Lee
Minneapolis-Honeywell
Minneapolis, Minn.

[3] F. B. Smith, "Time-optimal control of higher-order systems," IRE Trans. on Automatic Control, vol. AC-6, pp. 16–21; February, 1961.

Comments on "Mathematical Aspects of the Synthesis of Linear Minimum Response-Time Controllers"*

In a recent paper by Lee,[1] use was made of the Lagrange multipliers for solving the time-optimal control problem. The solution procedure as stated by Lee consisted of minimizing the total response time $t_r = F(t_1 \cdots t_k)$ [12][2] subject to the following constraints:

$$G_i(t_1 \cdots t_k) = 0, \quad i = 1, 2, \cdots n - 1, \quad (1)$$

and

$$0 \leq t_1 \leq t_2 \leq \cdots \leq t_k \leq t_r. \quad (2)$$

In the discussion following the paper, Hopkin stated that "The first constraint is taken care of by [14] (Lagrange multipliers), but the second (referring to the inequalities) is not."

The purpose of this note is to indicate that the inequality constraints can also be taken care of by Lagrange multipliers by noting that (2) may be written as K inequalities:

$$t_1 \geq 0, \quad t_2 - t_1 \geq 0,$$
$$t_3 - t_2 - t_1 \geq 0, \cdots, t_r - t_k \cdots - t_1 \geq 0. \quad (3)$$

For mathematical convenience, a set of new variables y_i[2] is introduced so that the inequalities may be written as equalities:

$$t_1 = y_1^2, \quad t_2 - t_1 = y_2^2,$$
$$t_3 - t_2 - t_1 = y_3^2, \cdots, t_k - t_{k-1} \cdots t_1 = y_{k-1}^2$$

and

$$t_r - t_k \cdots - t_1 = F(t_1 \cdots t_k) - t_k \cdots - t_1 = y_k^2. \quad (4)$$

The modified function corresponding to [14] has the form:

$$f' = F(t_1 \cdots t_k) + \sum_{i=1}^{n-1} \lambda_i G_i(t_1 \cdots t_k)$$
$$+ \lambda_n(t_1 - y_1^2) + \lambda_{n+1}(t_2 - t_1 - y_2^2) \cdots$$
$$+ \lambda_{n-1+k-1}(t_k - t_{k-1} \cdots - t_1 - y_{k-1}^2)$$
$$+ \lambda_{n-1+k}[F(t_1 \cdots t_k) - t_k \cdots - t_1 - y_k^2].$$

* Received by the PGAC, December 12, 1960.
[1] E. B. Lee, IRE Trans. on Automatic Control, vol. AC-5, pp. 283–290; September, 1960.

Received by the PGAC, January 3, 1961.

* Received by the PGAC, March 23, 1961.
[1] E. B. Lee, IRE Trans. on Automatic Control, vol. AC-5, pp. 283–290; September, 1960.
[2] [ ] and ( ) will be used to denote equations in Lee's paper and in this note, respectively.

The necessary conditions for an extremal become

$$\frac{\partial f'}{\partial t_i} = 0; \quad \frac{\partial f'}{\partial y_i} = 0; \quad i = 1, 2 \cdots k; \tag{6}$$

or:

$$\left.\begin{aligned} &\frac{\partial F}{\partial t_1} + \sum_{i=1}^{n-1} \lambda_i \frac{\partial G_i}{\partial t_i} + \lambda_n - \lambda_{n+1} - \lambda_{n+2} \cdots + \lambda_{n-1+k}\left[\frac{\partial F}{\partial t_1} - 1\right] = 0, \\ &\qquad\vdots \qquad\qquad\qquad\qquad \vdots \\ &\frac{\partial F}{\partial t_k} + \sum_{i=1}^{n-1} \lambda_i \frac{\partial G_i}{\partial t_k} + \lambda_{n-1+k-1} + \lambda_{n-1+k}\left[\frac{\partial F}{\partial t_k} - 1\right] = 0; \end{aligned}\right\} \tag{7a}$$

$$\left.\begin{aligned} &\frac{\partial F}{\partial y_1} = -2\lambda_n y_1 = 0, \\ &\frac{\partial F}{\partial y_2} = -2\lambda_{n-1} y_2 = 0, \\ &\qquad\vdots \qquad\qquad \vdots \\ &\frac{\partial F}{\partial y_k} = -2\lambda_{n-1+k} y_k = 0. \end{aligned}\right\} \tag{7b}$$

Eq. (7b) implies that either y_i or λ_{n-1-i} must be zero. However, (4) implies that if y_i vanishes, then $t_{i+1} - t_i \cdots - t_1$ also vanishes.

These conditions in turn imply that

$$\left.\begin{aligned} &\lambda_{n-1+k}[F(t_1 \cdots t_k) - t_k - t_{k-1} \cdots - t_1] = 0, \\ &\qquad\vdots \\ &\lambda_{n+1}(t_2 - t_1) = 0, \\ &\lambda_n t_1 = 0. \end{aligned}\right\} \tag{8}$$

Moreover, it can be shown that an additional necessary condition for minimum t_r is:

$$\lambda_i \leq 0, \quad i = 1, 2 \cdots, n - 1 - k. \tag{9}$$

Thus, the necessary conditions for minimum response time consist of (7a) and (8) and inequality sets (3) and (9). However, it is difficult to derive general necessary and sufficient conditions for *global minimum* over the constraint set defined by (3). The causes of the complications are:

1) The function to be minimized, *i.e.*, $t_r = F(t_1 \cdots t_k)$, enters into the last inequality constraint.
2) The behavior of $F(t_1 \cdots t_k)$ within a closed region (defined by the constraints) in $t_1, t_2 \cdots t_k$ space may be quite complex, as stated by Hopkin.

For real-time control of many high-order systems with slowly time-varying inputs, the forcing polarity during the first switching-time interval $(t, t+t_1)$ is of primary importance. In this case, the first complication may be removed by neglecting the last inequality, thus reducing the inequality constraints to a set of linear inequalities.

If the system state is sufficiently close to the desired state to permit analytical approximations of the time-domain trajectory equations (*e.g.*, polynomial approximation), then the global behavior of t_r within the constraint region can usually be deduced. This permits rough estimation of the optimum switching times.

The time-optimal control problem formulated by Lee reduces to a complex nonlinear programming problem. At the present time, rapidly converging numerical procedures for solving this type of problem on present-day high-speed digital computers, with sufficiently short computing time for real-time control, are still lacking. Further developments in this direction would help to close the gap between theoretical results and physical implementation of optimum control systems.

P. K. C. Wang
Res. Lab.
IBM Corp.
San Jose, Calif.

Comment on "An Optimal Strategy for a Saturating Sampled-Data System"*

In their paper,[1] Desoer and Wing have proved some very useful and interesting theorems for finding and implementing an optimal strategy for a special type of sampled-data system. However, the authors seemed to have lost rigor in stating and proving Theorem 1 and Theorem 2. The terms "regions," "convex regions" and "closed sets," which have precise meanings, have been loosely used. For example, in Theorem 1, R_N', which is defined as a set of initial states that can be brought to equilibrium in N sampling periods or less, is stated as a region. Since not every set is a region, the statement of Theorem 1 should have been backed by a formal proof.

The following may be substituted in place of Theorem 1 and Theorem 2 (pp. 7–8):

Theorem 1: The set R_N' is convex, *i.e.*, if two initial states represented by the points P_1 and P_2 can be brought to equilibrium in N sampling periods or less, the same is true for any initial state on the line segment P_1P_2.

Proof: As in the above-mentioned paper.

Let us recall the definition of a convex region before stating Theorem 2.

Definition: A convex set (on a plane) which is closed and contains at least one interior point (an interior point is a point of the set which is not a boundary point) is called a convex region.

It immediately follows from the above definition that the set R_1' which is a line segment (closed and convex), is not a convex region since it does not contain an interior point.

Theorem 2: The convex set $R_N'(N>1)$ is a bounded convex region whose boundary is formed by the edges of a polygon Π_N which has the following $2N$ vertices:

$$\begin{aligned} OP_1 &= r_1 - r_2 - \cdots - r_N \\ OP_2 &= r_1 + r_2 - \cdots - r_N \\ &\vdots \\ OP_N &= r_1 + r_2 + \cdots + r_N \\ OP_{-1} &= -r_1 + r_2 + \cdots + r_N \\ OP_{-2} &= -r_1 - r_2 + \cdots + r_N \\ &\vdots \\ OP_{-N} &= -r_1 - r_2 - \cdots - r_N, \end{aligned}$$

where O is the origin of the (γ_1, γ_2) plane.

Outline of the Proof: From the definition of the vectors $r_k (k = 1, \cdots, N)$, it is not difficult to show that the polygon Π_N (with its edges and interior) is an intersection of a finite number of convex regions and whence follows the convexity of the polygon Π_N. By the famous Jordan curve theorem, the edges of the polygon Π_N divides the (γ_1, γ_2) plane into two regions—an inside and an outside. Let P be an arbitrary interior point of Π_N. (Such a point exists since Π_N is a convex region.) From the convexity of Π_N, it follows that a ray (half line) from the origin through P intersects the boundary of Π_N in exactly one point. Let that point be Q_k. The rest of the proof could be constructed, *mutatis mutandis* from the proof of Theorem 2 by Desoer and Wing.

I. H. Mufti
Analysis Section
Natl. Res. Council
Ottawa, Ont., Canada

Authors' Reply[2]

We agree with Mr. Mufti that the word region should not have been used in the paper,[1] especially in the statement of theorem 1. The damage, however, is very slight since all R_N', for $N>1$, are actually regions and R_1' is not, although it is a closed convex

* Received by the PGAC, April 23, 1961.

[1] C. A. Desoer and J. Wing, IRE Trans. on Automatic Control, vol. AC-6, pp. 5–15; February, 1961.

[2] Received by the PGAC, May 4, 1961.

set. In all results following theorem 1, only the fact that R_N' is a closed convex set, for $N=1, 2, \cdots$, is used; hence no result, except theorem 1, is invalidated by this defect in terminology. We suggest that the reader replace the word "region" at the six places where it occurs by the word "set."

Mr. Mufti rewords our theorem 2 in order to stress the fact that one should prove that the polygon Π_N is convex and he proves this fact by the Jordan Curve Theorem. We agree with him and thank him for supplying such a proof. His proof cannot, however, be generalized easily to an n-dimensional state space because of its use of the Jordan Curve Theorem. In a paper[3] already submitted for publication, the authors have generalized the results of this paper for nth-order systems whose transfer function has real distinct nonpositive poles. In particular, they have shown that: R_N' is the convex hull of the set V_N where V_N is the set of all points of the form

$$\sum_{i=1}^{N} \epsilon_i r_i$$

with $\epsilon_i = \pm 1$ and the sequence ϵ_i has no more than $n-1$ sign variations. This is a generalization to nth order systems of theorem 2 of the paper under discussion.

C. A. Desoer
J. Wing
Elec. Engrg. Dept.
University of California
Berkeley, Calif.

[3] C. A. Desoer and J. Wing, "A minimal time discrete system," IRE Trans. on Automatic Control, vol. AC-6, pp. 111–125; May, 1961. Also Electronics Res. Lab., Univ. of California, Berkeley, Ser. No. 60, Issue 327; November 15, 1960.

Correction

Mr. C. A. Desoer has called the following to the attention of the *Editor*. On page 8, in the first column of this paper,[1] line 5 from the top should read:

"Let P be an arbitrary interior point of the complex polygon Π_N."

On the Time-Optimal Regulation of Plants with Numerator Dynamics*

Determination of the time-optimal control for plants described by a differential equation in which derivatives of the control variable appear is a difficult problem. In most engineering problems of this type, the time-optimal control is a relay-type control. When this relay control is used on a plant with numerator dynamics, some of the variables used to describe the movement become discontinuous. For this reason it is desirable to consider a coordinate system in which no derivatives of the control variable appear. Such a coordinate system can be found using the procedure described by Laning and Battin.[1] We must then answer the question: after such a transformation, what is the definition of our original time-optimal control problem and how does one proceed to solve it? We show that the problem statement remains basically the same, and that Pontriagin's Maximum Principle can be used in defining the form of the control. We reduce the regulation problem to that of steering the state vector from the initial state to a region in the state space, and show that the time-optimal control is a relay-type control.

Consider a plant described by the differential equation

$$\frac{d^n x}{dt^n} + a_1 \frac{d^{n-1}x}{dt^{n-1}} + \cdots + a_{n-1}\frac{dx}{dt} + a_n x = b_0 \frac{d^m y}{dt^m} + b_1 \frac{d^{m-1}y}{dt^{m-1}} + \cdots + b_m y, \quad (1)$$

where $m<n$, a_i, b_j = const., and the control variable y is subject to saturation, *i.e.*, $|y| \leq 1$.

The time-optimal regulation problem is to take $x(t)$ from a given initial condition x_0' to the origin $x=0$ in minimum time t_r with $|y| \leq 1$, and such that $x(t)=0$ for $t \geq t_r$ if no additional disturbances are applied. This implies bringing $x(t)$ and its first $n-1$ derivatives to zero at the time t_r.

Using the transformation given by Laning and Battin,[1] (1) becomes

$$x = x_1$$

and

$$\begin{aligned}
\dot{x}_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + b_{11}y \\
\dot{x}_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + b_{22}y \\
&\vdots \\
\dot{x}_n &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n + b_{nn}y, \qquad |y| \leq 1, \quad (2)
\end{aligned}$$

in which no derivatives of the control variable y appear. If we require that $x_1(t_r)=0$ and that $x_1(t)=0$ for $t \geq t_r$ in minimum time t_r with $|y|<1$, the problem is the same as the original one.

The condition that $x_1(t)=0$ for $t \geq t_r$ implies that $\dot{x}_1(t)=0$ for $t>t_r$. This condition in turn implies that

$$a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n + b_{11}y = 0 \quad \text{for } t \geq t_r. \quad (3)$$

Hence, the original problem is the same as the state vector control problem in which one of the variables $x_i(t)$ is to be brought to zero and held there by means of a control, $|y| \leq 1$.

Assume for the remaining discussion that the state variable which is to be controlled is $x_1(t)$ and that the order of the control variable derivatives in (1) is $m=n-1$. In the other cases, it will be obvious how to proceed.

From (3), for $\dot{x}_1(t)=0$ with $t \geq t_r$, we will construct a region G in the n-dimensional state space R^n such that with $|y| \leq 1$ we can keep $x_1(t)=0$ for all $t \geq t_r$. To do this, we note that the region G is exactly the set of values that $x_1(t)=0$, $x_2(t)$, $x_3(t) \cdots x_n(t)$ can have at time $t=t_r$, such that with $|y| \leq 1$, (3) will be satisfied for all $t \geq t_r$. Solving (3) for the $y(t)$, which we must use after getting to G at time $t=t_r$ in order to keep $x(t)=0$ for $t \geq t_r$, gives

$$y(t) = -\left[\frac{a_{12}x_2(t)}{b_{11}} + \frac{a_{13}x_3(t)}{b_{11}} + \cdots + \frac{a_{1n}x_n(t)}{b_{11}}\right] \quad (4)$$

with $|y| \leq 1$, $t \geq t_r$.

This $y(t)$ is substituted into (2) to obtain the homogeneous differential equation

$$\begin{aligned}
\dot{x}_2 &= \alpha_{22}x_2 + \alpha_{23}x_3 + \cdots + \alpha_{2n}x_n \\
\dot{x}_3 &= \alpha_{32}x_2 + \alpha_{33}x_3 + \cdots + \alpha_{3n}x_n \\
&\vdots \\
\dot{x}_n &= \alpha_{n2}x_2 + \alpha_{n3}x_3 + \cdots + \alpha_{nn}x_n. \quad (5)
\end{aligned}$$

The solution of (5) can be written as

$$\begin{aligned}
x_2(t) &= \phi_{22}(t)x_{20} + \phi_{23}(t)x_{30} + \cdots + \phi_{2n}(t)x_{n0} \\
x_3(t) &= \phi_{32}(t)x_{20} + \phi_{33}(t)x_{30} + \cdots + \phi_{3n}(t)x_{n0} \\
&\vdots \\
x_n(t) &= \phi_{n2}(t)x_{20} + \phi_{n3}(t)x_{30} + \cdots + \phi_{nn}(t)x_{n0}, \quad (6)
\end{aligned}$$

where $x_{i0}=x_i(t_r)$; $i=2, 3 \cdots n$.

If this result is substituted into (4), we obtain a condition of the form

$$\left|\beta_2(t)x_{20} + \beta_3(t)x_{30} + \cdots + \beta_n(t)x_{n0}\right| \leq 1 \quad (7)$$

for all $t \geq t_r$. The region G is the values of x_{i0} for which it is possible with $|y| \leq 1$ to stay on the hyperplane $x_1=0$, which is exactly the condition (7). Hence, the region G can be constructed as follows: Let $J(t_1)$ be the values of $x_{20}, x_{30} \cdots x_{n0}$ such that

$$\left|\beta_2(t_1)x_{20} + \beta_3(t_1)x_{30} \cdots \beta_n(t_1)x_{n0}\right| \leq 1$$

[written

$$J(t_1) = \{x_{20}, x_{30} \cdots x_{n0} \mid \left|\beta_2(t_1)x_{20} + \beta_3(t_1)x_{30} \cdots + \beta_n(t_1)x_{n0}\right| \leq 1\}].$$

Consider the intersection of the sets $J(t_1)$ for various t_1 in the interval (t_r, ∞) written

$$J = \bigcap_{t_1 \in (t_r, \infty)} J(t_1).$$

J is the values of $(x_{20}, x_{30} \cdots x_{n0})$ such that (7) is satisfied for all $t \geq t_r$. From J, consider only those values of $(x_{20}, x_{30} \cdots x_{n0})$ which are in the hyperplane $x_1=0$; this is the region G.

What properties does G have? It is a closed region being the intersection of many closed regions of the form

$$\left|\beta_2(t_1)x_{20} + \beta_3(t_1)x_{30} \cdots + \beta_n(t_1)x_{n0}\right| \leq 1.$$

It is convex because any two of its points can be joined by a straight line segment which has all of its points in G. It is nonempty because the point $x_{10}=0$, $x_{20}=0 \cdots x_{n0}=0$ is certainly in it.

The time-optimal single-variable state vector control problem is, in terms of the region G, to bring the plant from some arbitrary initial state $x_{10}', x_{20}' \cdots x_{n0}'$ to an

* Received by the PGAC, March 23, 1961.

[1] J. H. Laning and R. H. Battin, "Random Processes in Automatic Control," McGraw-Hill Book Co., Inc., New York, N. Y., p. 191; 1956.

intersection with the region G in minimum time t_r with $|y| \leq 1$.

According to Rozonoer,[2] if the region G is closed and convex, Pontriagin's Maximum Principle is a necessary condition which the optimum control must satisfy. For our linear problem, the maximum principle allows the optimum control to be at a boundary point (usually relay control) until we intersect the region G. After intersecting G, y must be changed so as to keep $(x_1(t), x_2(t) \cdots x_n(t))$ in G for all $t \geq t_r$. For $(x_1(t), x_2(t) \cdots x_n(t))$ in G, y can be found from (3).

The construction of the optimum control to get to the region G is very difficult and has been solved in only a number of special cases (see, for example, Schmidt[3]). The transversality conditions as given by Pontriagin[4] help to resolve the question of which point in G we should aim for. If it can be reduced to a small number of points, it is possible to solve this problem by using switching equations.[5] If $m=1$ in (1), then G is a line segment, and by assigning a parameter to this segment it is often possible to solve for the optimum control.[6]

E. B. Lee
Minneapolis-Honeywell Reg. Co.
and University of Minnesota
Minneapolis, Minn.

[2] L. I. Rozonoer, "L. S. Pontriagin's principle of maximum in the theory of optimum systems," *Avtomati. i Telemekh.*, vol. 20, pp. 1441–1458; 1959.
[3] S. Schmidt, "The Analysis and Design of Continuous and Sampled Data Feedback Control Systems with a Saturation Type Nonlinearity," Ph.D. dissertation, Stanford University, Stanford, Calif.; 1959.
[4] V. G. Boltanskii, R. V. Gamkrelidze, and L. S. Pontriagin, "The theory of optimal processes (I—The Maximum Principle)," *Izvestia Akad. Nauk SSSR, Ser. Mat.*, vol. 24, pp. 3–42; 1960.
[5] E. B. Lee, "Mathematical aspects of the synthesis of linear minimum response time controllers," IRE Trans. on Automatic Control, vol. AC-5, pp. 283–289; September, 1960.
[6] C. A. Harvey, "On determining the switching criterion for time-optimal control," to be published.

On Third-Order Time-Optimal Control Systems*

The system under consideration consists of a linear dynamic process describable by

$$\frac{d\bar{C}}{d\tau} = \Omega\bar{C} + \bar{F}(\tau), \tag{1}$$

where Ω and $\bar{C}(\tau)$ are the constant square matrix and process state vector, respectively. $\bar{F}(\tau)$ is a vector forcing function with limiting constraint imposed on its components $f_i(\tau)$:

$$|f_i(\tau)| \leq K_i, \quad i = 1, 2, 3, \cdots N. \tag{2}$$

The control problem is to find the required $\bar{F}(\tau)$ such that $\bar{C}(\tau)$ coincides with a time-varying input vector $\bar{R}(\tau)$ in minimum time.

* Received by the PGAC, March 29, 1961.

It has been shown that for normal processes,[1,2] the time-optimal forcing components $f_i(\tau)$ take on their extreme values. This result justifies the well-known procedures for determining the optimum switching boundaries in phase space. However, in the above case, the switching boundaries vary with the input $\bar{R}(\tau)$.

In a recent paper,[3] approximate analytical expressions for the optimum switching surfaces of third-order processes have been derived by extrapolating both the input and the process time-domain trajectories in the form of Taylor series. However, no results are given concerning the accuracy of the approximate switching surface and its effect on the over-all system response. The purpose of this correspondence is to examine these points and to clarify some of the obscure points in the derivation of the switching criteria.

Using the vector notation, the future input state is extrapolated by

$$\bar{R}(t+\tau) = \begin{bmatrix} r(t+\tau) \\ \dot{r}(t+\tau) \\ \ddot{r}(t+\tau) \end{bmatrix} = \begin{bmatrix} 1 & \tau & \tau^2/2 \\ 0 & 1 & \tau \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r(t) \\ \dot{r}(t) \\ \ddot{r}(t) \end{bmatrix} + \dddot{r}(t) \begin{bmatrix} \tau^3/3! \\ \tau^2/2 \\ \tau \end{bmatrix}, \tag{3}$$

where t and τ are the present and future times, respectively.

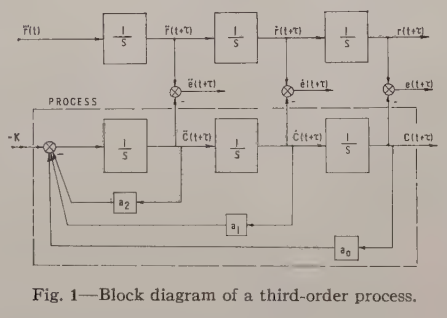


Fig. 1—Block diagram of a third-order process.

For a process (Fig. 1) describable by

$$\sum_{n=0}^{3} a_n \frac{d^n c}{d\tau^n} = \pm K \quad \text{with} \quad a_3 = 1,$$

Ω and $\bar{F}(\tau)$ have the forms

$$\Omega = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 \end{bmatrix};$$

$$\bar{F}(\tau) = \pm K \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \pm K\bar{F}_0.$$

[1] R. V. Gamkrelidze, "The theory of optimum speed of response in linear systems," *Izvestia Akad. Nauk SSSR, Ser. Mat.*, vol. 22, pp. 449–474; July–August, 1958.
[2] J. P. LaSalle, "Time optimal control systems," *Proc. Natl. Acad. Sci.*, vol. 45, pp. 573–577; April, 1959.
[3] A. M. Hopkin and P. K. C. Wang, "Further work on relay type control systems designed for random inputs," *Proc. IFAC*, Congress, Butterworth Scientific Publications, London, England; July, 1960.

The process time-domain trajectories corresponding to N polarity reversals of forcing K are found by repeated applications of the solution to (1):

$$\bar{C}(t+\tau) = \epsilon^{\Omega\tau}\bar{C}(t) \pm K \sum_{n=1}^{N} (-1)^{n-1} \cdot \int_0^{t_n} \epsilon^{\Omega(\tau - t_{n-1} - t')}\bar{F}_0 dt', \tag{4}$$

where

$$\tau = \sum_{n=1}^{N} t_n \quad \text{and} \quad t_0 = 0.$$

For third-order processes, the case with $N=2$ is of particular interest (for oscillatory processes, restrictions must be placed on the initial states so that only two switchings are necessary). An approximate extrapolation of the error state vector $\bar{E}$ is obtainable by expanding the fundamental matrix $\epsilon^{\Omega\tau}$ as a truncated power series in time and then combining (3) and (4)

$$\bar{E}(t+\tau) = \bar{R}(t+\tau) - \bar{C}(t+\tau) = \begin{bmatrix} e(t+\tau) \\ \dot{e}(t+\tau) \\ \ddot{e}(t+\tau) \end{bmatrix} = \begin{bmatrix} 1 & \tau & \tau^2/2 \\ 0 & 1 & \tau \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} e(t) \\ \dot{e}(t) \\ \ddot{e}(t) \end{bmatrix} + \delta^{\pm} \begin{bmatrix} \tau^3/3! \\ \tau^2/2 \\ \tau \end{bmatrix} \pm K \begin{bmatrix} t_2^3/3 \\ t_2^2 \\ 2t_2 \end{bmatrix}, \tag{5}$$

where $\tau = t_1 + t_2$ and

$$\delta^{\pm} = \dddot{r}(t) \mp K + \sum_{n=0}^{2} a_n [r^{(n)}(t) - e^{(n)}(t)]. \tag{6}$$

The $+$ and $-$ signs on δ denote the polarity of K during the time interval $(t, t+t_1)$. Clearly, the approximation becomes more accurate as the coefficients $a_n \to 0$. For a triple-integral process ($a_n = 0$), (5) gives exact extrapolation of E.

Setting $\bar{E}(t+\tau)$ to zero and eliminating t_1 and t_2 leads to approximate expressions of the switching boundaries in terms of $e(t)$, $\dot{e}(t)$ and $\ddot{e}(t)$ only.

For $+K$:

$$e^{+}(t) = \frac{\ddot{e}(t)}{\delta^{+}}\left[\dot{e}(t) - \frac{\ddot{e}^2(t)}{3\delta^{+}}\right] - \frac{(4K+\delta^{+})}{6(\delta^{+})^2}\left\{\frac{[\ddot{e}^2(t) - 2\dot{e}(t)\delta^{+}]^3}{2K(2K+\delta^{+})}\right\}^{1/2}, \tag{7}$$

and for $-K$:

$$e^{-}(t) = \frac{\ddot{e}(t)}{\delta^{-}}\left[\dot{e}(t) - \frac{\ddot{e}^2(t)}{3\delta^{-}}\right] - \frac{(4K-\delta^{-})}{6(\delta^{-})^2}\left\{\frac{[\ddot{e}^2(t) - 2\dot{e}(t)\delta^{-}]^3}{-2K(-2K-\delta^{-})}\right\}^{1/2}. \tag{8}$$

The intersection between $e^{+}(t)$ and $e^{-}(t)$ generates a curve representing a discontinuity in the switching surface. It can be easily verified by using the obvious relation $\delta^{-} = \delta^{+} + 2K$ that this curve constituted by the terminal trajectories corresponding to forcing K and $-K$. The expressions for time t_2 and the terminal trajectories are obtain-

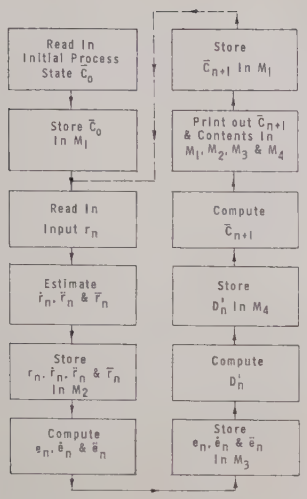


Fig. 2—Computing cycle for the simulated system.

able by setting $t_1=0$ in (5)

$$t_2 = \frac{-\ddot{e}(t)}{\delta^{\pm} \pm 2K} \equiv \frac{-\ddot{e}(t)}{\delta^{\mp}}; \tag{9}$$

$$\dot{e}(t) = \frac{\ddot{e}^2(t)}{2(\delta^{\pm} \pm 2K)} \equiv \frac{\ddot{e}^2(t)}{2\delta^{\mp}}. \tag{10}$$

The terminal trajectories can be also expressed as

$$e(t) = \frac{\dddot{e}^3(t)}{6\delta^{\mp}}. \tag{11}$$

If the input is slowly time-varying so that

$$\left| \dddot{r}(t) + \sum_{n=0}^{2} a_n[r^{(n)}(t) - e^{(n)}(t)] \right| < K, \tag{12}$$

then $\delta^+<0$ and $\delta^->0$. Consequently, the following relation holds for positive t_2:

$$-\operatorname{sgn} \delta^{\mp} = \operatorname{sgn} \ddot{e}(t). \tag{13}$$

Hence, (10) and (11) may be rewritten as

$$\dot{e}(t) = \frac{\ddot{e}^2(t)}{2\left\{\dddot{r}(t) - K \operatorname{sgn} \ddot{e}(t) + \sum_{n=0}^{2} a_n[r^{(n)}(t) - e^{(n)}(t)]\right\}}; \tag{14}$$

$$\dddot{e}(t) = \frac{\dddot{e}^3(t)}{6\left\{\dddot{r}(t) - K \operatorname{sgn} \ddot{e}(t) + \sum_{n=0}^{2} a_n[r^{(n)}(t) - e^{(n)}(t)]\right\}}. \tag{15}$$

For $a_0=0$, an error deviation $D(t)$ may be used for switching decision:

$$D(t) = e(t) - e^{\pm}[\dot{e}(t), \ddot{e}(t)]. \tag{16}$$

However, it is more convenient to measure the error deviation in a translated coordinate system whose origin is at the discontinuity of the switching surface projected onto the plane $[e(t), \dot{e}(t), \ddot{e}(t)=0]$; *i.e.*,

$$D'(t) = [e(t) - e_t(t)] - e^{\pm}[\dot{e}(t) - \dot{e}_t(t), \ddot{e}(t)], \tag{17}$$

where $e_t(t)$ and $\dot{e}_t(t)$ are the error and error rate on the terminal trajectories, which are obtainable from (11) and (10), respectively.

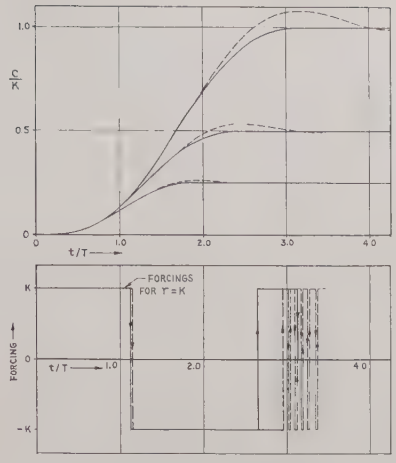


Fig. 3—Step response of a two-integral plus one time-constant third-order process. Solid lines: optimum response; dashed lines: near optimum response; $\Delta t = 0.025T$.

The final form of the forcing vector $F(t)$ is

$$\bar{F}(t) = K \operatorname{sgn} D'(t)\bar{F}_0. \tag{18}$$

The above approximate switching criterion is useful when the system trajectories are restricted to a finite region about the origin of error phase space. The region size is determined by the total anticipated transient time and the maximum permissible error in approximation. In the case where Ω has complex eigenvalues, the useful region can be determined by restricting the total transient within $\frac{1}{2}$ cycle of the oscillatory response.

A closed-loop system with a two-integral plus one time-constant (T) process ($a_2=1/T$, $a_1=a_0=0$) was simulated on an IBM 650 digital computer. The input and process state variables were sampled periodically with period Δt. The computing cycle is given in Fig. 2. The exact process response ($\bar{C}_{n+1}$) at the $n+1$-th sampling instant subjecting to $\bar{F}(t)$(18) was determined from the state transition equation

$$\bar{C}_{n+1} = \Lambda\bar{C}_n + K \operatorname{sgn} D_n'\bar{\Gamma}, \tag{19}$$

where

$$\bar{C}_n = \begin{bmatrix} c_n \\ \dot{c}_n \\ \ddot{c}_n \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 1 & \Delta t & T[\Delta t - T(1-\epsilon^{-\Delta t/T})] \\ 0 & 1 & T(1-\epsilon^{-\Delta t/T}) \\ 0 & 0 & \epsilon^{-\Delta t/T} \end{bmatrix}, \quad \bar{\Gamma} = \begin{bmatrix} \Delta t^2/2 - T\Delta t - T^2(1-\epsilon^{-\Delta t/T}) \\ \Delta t - T(1-\epsilon^{-\Delta t/T}) \\ 1-\epsilon^{-\Delta t/T} \end{bmatrix},$$

and

$$D_n' = e_n - e_{tn} - \frac{\ddot{e}_n}{\delta_r}\left[\dot{e}_n - \dot{e}_{tn} - \frac{\ddot{e}_n^2}{3\delta_n}\right] - \frac{[4K \operatorname{sgn}(\dot{e}_n - \dot{e}_{tn} - \delta_n)]}{6\delta_n^2}\left\{\frac{[\ddot{e}_n^2 - 2\delta_n(\dot{e}_n - \dot{e}_{tn})]^3}{2K[2K - \delta_n \operatorname{sgn}(\dot{e}_n - \dot{e}_{tn})]}\right\}^{1/2},$$

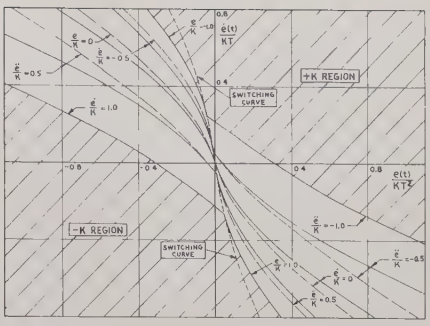


Fig. 4—Exact and approximate switching boundaries (denoted by solid and dashed lines, respectively) for step and ramp inputs.

with

$$e_{tn} = \frac{\ddot{e}_n^3}{6\left[\dddot{r}_n - K \operatorname{sgn} \ddot{e}_n - \frac{1}{T}(\ddot{r}_n - \ddot{e}_n)\right]},$$

$$\dot{e}_{tn} = -\frac{\ddot{e}_n^2}{2\left[\dddot{r}_n - K \operatorname{sgn} \ddot{e}_n - \frac{1}{T}(\ddot{r}_n - \ddot{e}_n)\right]},$$

$$\delta_n = \dddot{r}_n - K \operatorname{sgn}(\dot{e}_n - \dot{e}_{tn}) + \frac{1}{T}(r_n - \ddot{e}_n).$$

The system response for various step position inputs is shown in Fig. 3. Results indicate that the response to a step position input with magnitude K has approximately 7 per cent overshoot. The percentage of overshoot becomes progressively less as the input magnitude becomes smaller. This is tolerable for most practical applications. The steady-state oscillations are due to the sampling nature of the simulated system. Fig. 4 shows the exact and approximate switching surfaces for step and ramp inputs. It can be seen that the use of the approximate surface in the switching criteria has the effect of a small switching time delay. This effect can be compensated by introducing a small correction signal in D_n'. Tests were also made using random inputs generated by a Gaussian noise source. The noise was filtered and sampled periodically by means of a zero-order hold circuit and a digital voltmeter. Random input sequences were recorded and fed into the computer. The first- and second-order derivatives were estimated by taking first and second differences, respectively. $\dddot{r}(t)$ was assumed to be zero. Test results are similar to those reported previously by Hopkin and Wang.[3] However, the positional error-amplitude probability distribution varies with the sampling period Δt. In the control of physical processes with on-line digital computers,

Δt is governed primarily by the required computation time or the speed of the available computer. Small Δt generally gives "tighter" closed-loop control of the processes.

It has been shown in this correspondence that a noniterative, near time-optimal controller is feasible for a third-order process with bounded trajectory variations and slowly time-varying inputs. In a recent paper by Smith,[4] an iterative time-optimal digital controller for a higher-order process was successfully demonstrated for small input variations. Further work in developing rapidly converging iterative time-optimal digital controllers would have significant practical value.

P. K. C. Wang
Res. Lab.
IBM Corp.
San Jose, Calif.

[4] F. B. Smith, Jr., "Time-Optimal Control of Higher-Order Systems," presented at IRE Regional Conf., Seattle, Wash.; May, 1960.

Controlled Camping or USSR in Heterospect*

Visa

I was scheduled to present a paper to the First Congress of the International Federation on Automatic Control to be held in Moscow, June 27 to July 6, 1960. As Chairman of the Education Committee of the American Automatic Control Council, I was also a member of the Education Committee of the International Federation on Automatic Control, which was to meet on June 24 in Moscow. We had heard of numerous cases in which Russian visas were delayed until after the boat sailed, or even until after the Congress had been held. We therefore resolved to arrive in Russia significantly in advance of the Congress. In the spring of 1959, I wrote to the authors of many papers in *Avtomatika i Telemekhanika* and received a few replies. From these contacts, I planned to visit Odessa, Kiev Kharkov, Yalta, Moscow, and Leningrad. To have freedom of motion, it was essential for me to take my own automobile, and this led to the decision to take my wife, my daughter, aged 16, and my sons, aged 14, 12, and 10.

In the fall of 1959, we applied to Intourist for permission to take a 60-day tour in our own car, starting at Istanbul and going by Russian boat to Yalta, and then driving to Kharkov, west to Kiev, south to Odessa, and back north to Kiev, east to Kharkov, north to Moscow, and on to Leningrad and Finland. This was refused because the road from Odessa to Kiev was closed to tourists, and the boat from Istanbul went to Odessa instead of to Yalta. Negotiations continued for many months and were strengthened by invitations to me to give lectures at institutes and to stay in the homes of professors in Russia.

I insisted on no Intourist guide and no planned excursions. Eventually we were granted permission for a 40-day trip going by boat to Odessa, and then driving 700 km north to Kiev, 500 km east to Kharkov, 1000 km south to Yalta, 100 km north to Simferopol, 300 km north to Melitopol, 150 km north to Zaporoshe, 330 km north to Kharkov, 230 km north to Kursk, 180 km north to Orjel, 400 km north to Moscow, 200 km northwest to Kalinin, 400 km northwest to Novgorod, and 200 km northwest to Leningrad.

On the usual Intourist tours, one must stay in a prescribed Intourist hotel each night, and is therefore under serveillance, with an inflexible schedule. On the usual Intourist tours, one's hotel and meal expenses in Russia are prepaid at the rate of 4 rubles per dollar.

On camping tours, one can increase or decrease his length of stay in any city easily, although an extension of the total time is impossible. The quality of accommodations can always be upgraded. One can eat at whatever restaurant he wishes and can stay in a hotel if he so desires. He can rent a car in Helsinki or Leningrad. On camping tours, one pays one dollar per night per person for his visa, and then pays his expenses inside of Russia with cash at 10 rubles per dollar. I therefore requested a camping tour, taking my own Volkswagen Microbus, tent, and sleeping bags.

After my target date for leaving Germany had passed without a firm agreement with Intourist, I filed an application for a Russian visa with the embassy in Bonn. It was eventually necessary to telephone to Intourist in Moscow for final approval of a 40-day trip without an Intourist guide, with camping accommodations for the family in all cities, and with an extra hotel room for me in Moscow and Leningrad.

Armed with an Intourist receipt for my dollars, I was able to get a visa from Bonn in one day instead of the usual one or two weeks, because we had previously filed an application which had been tentatively approved. We should have had a special Russian visa for our automobile listing the license number, but this we did not know about. Although we had been assured by Moscow of the boat reservations from Istanbul to Odessa, my periodic letters to the boat company in Odessa and in Istanbul from January through April had never been answered. In anticipation of difficulties, I had applied for transit visas for Bulgaria and Rumania, to the embassies in Bern in April. We drove to Bern and got these visas, which would permit us to enter Russia at Tchernovtsy on the border of Rumania.

Boat Trip

We bought $400.00 worth of canned goods, and with all of the children's books, my technical books and papers, extra gasoline, spare parts, and camping equipment, we drove to Trieste, Italy, then along the coast of Yugoslavia to Split and Dubrovnik, and inland to Skopje, the ancient capital of Macedonia. This latter was a particularly terrible road. Modern Greece was a relief, and we were very hospitably welcomed by the parents of Michael Athanassiades in Drama, Greece. We drove into Turkey when travel restrictions were lifted after the revolution.

In Istanbul I visited a college which had departments of chemistry, mathematics, physics, biology and medicine. It seemed to be well-equipped. I also visited Roberts College, which is well known as the American University for the Near East. The University of Istanbul was full of soldiers encamped on the lawns.

The day that our boat was to debark, we arrived at the docks to find that thirty-five more tickets had been sold than there were spaces on the Russian boat. We were, however, eventually accommodated. There was no one in charge of loading the car, and no slings were available. There was a debate whether it was more dangerous to drain the gas tank onto the dock, or to load the car with gasoline in the tank. The latter opinion prevailed, and it was put on deck with ropes slung under it, and with gasoline in its tank. Sailing out through the Bosporus in the evening, I went up on the navigational deck with the mate in charge, and took sightings with a theodolite and radar sightings to plot the course to Bulgaria. The radar was of Russian construction with approximately 1° accuracy. It had a problem of moding which was cured by momentarily turning off the power supply. It was used exclusively for intermittent sightings, and was not normally turned on.

Odessa

After two days we stopped in the Black Sea at noon before Odessa was in sight, and were boarded by a shipload of customs officers. They spent half a day searching minutely all of the baggage of the travelers. All money of all kinds and any item containing gold were carefully listed on the customs declaration. Every item of clothing in all of the suitcases and every piece of paper and book was carefully searched for hidden money or for propaganda material. Here we had difficulty getting clearance to remove the car because it was not listed in our passport. After we were cleared by customs, we bought rubles from the Intourist banker at 10 rubles per dollar.

In Odessa I had been invited to stay at the home of Professor A. S. Sadowskij. Professor Sadowskij hospitably met us on the boat in the evening when it docked and we sat down in the lounge to discuss engineering education. Another man joined us, however, and listened to our discussion for about an hour. He then volunteered to show us some of Odessa in his chauffeur-driven automobile. We saw the best streets and never got to the home of Professor Sadowskij and never met his wife and children. At 10 P.M. Professor Sadowskij was taken home by the other man, and at 11 P.M. we had our dinner in the Intourist hotel. We had a large room with six beds and a bath. At 12 P.M., the children retired and Phyllis and I returned to the boat to supervise the unloading of our car at 1:30 A.M. We met an Arab blackmarketeer who offered to sell us 25 rubles per dollar, which we declined. He helped us find the Intourist garage, which was the only safe place to leave the car at night. He called our attention to the large number of drunks and thieves.

* Received by the PGAC, February 7, 1961.

Sunday morning we picked up a small boy who spoke no English, but said he would help us find the gasoline station. This was not easy. He talked to dozens of truck drivers and after two hours we found the station on the east side of town. We took him home and started immediately for Kiev, using only a National Geographic Map which did not show the road we were taking. At Uman, half way to Kiev, we had to buy gasoline again. In this town we were directed over dirt streets to an empty lot in which stood a wooden shed two meters square. A teen-age neighbor boy disappeared on his bicycle and came back with the owner in about twenty minutes. By this time a large crowd had assembled and we had given a speech and passed out American pennies and our printed name and address. We flew our American flag and everyone said "peace and friendship." The plainclothes policeman who tried to object was told by the people to go mind his own business.

Gasoline was poured by hand from a fifty-gallon barrel into a three-gallon can, and from thence into our gas tank. After we had paid for the gasoline, a school teacher gave us a speech on behalf of the Russians. She said she wished more Americans would come to Russia and that the world could know that the Russian people did not want war. Some of the men who had suffered the loss of a limb said, "Never again."

On the road between Odessa and Kiev we saw a few trucks, but only two sedans during the entire day.

Kiev

We arrived in Kiev at midnight to find a welcoming letter left for us by Professor A. G. Ivakhnenko of the Institute of Electrotechniques of the Ukranian Academy of Sciences. We spent the next two days as his guests. Professor Ivakhnenko is the author of a textbook entitled "Technical Cybernetics" and has done considerable research on adaptive control systems. He has about six Ph.D. candidates working under his direction. Some were working on an optimum fuel-air ratio controller for a steam boiler. I saw a variable speed wound rotor induction motor, in which the rotor resistances were mounted in heat radiating plates on the shaft and the resistance values were centrifically controlled so that the machine would operate at maximum torque without overheating the windings at all values of speed from full-speed to the stalled condition. Professor Ivakhnenko had a digitally-controlled contour milling machine which read the input punched cards, calculated an interpolation schedule, and recorded a continuous control signal on magnetic tape. The tape, when fed into the milling machine, caused the work to move on a prescribed curve in two dimensions.

They had a well-equipped library with publications from all over the world in the field of control. His graduate students spoke English, and acted as translators when needed.

In other motor control problems, they were using negative-voltage feedback and positive-current feedback on an induction motor to obtain speed torque curves with flat sections at speeds other than near synchronous speed. They were using resistance proportional to slip in a three-phase wound rotor induction motor in order to reduce the heating at high slips. They were also working on magnetic amplifier controls, reversible ac motors, and transistor circuits.

They were working on a boiler feed-water resistivity controller to prevent corrosion. This minimum-corrosion controller was using a 10-minute period in a sampled system. The rate of addition of lime $Ca(OH)_2$ was changed after a 10-minute interval, and after a 20-minute interval the first difference in the resistivity was calculated to determine the optimum direction of average change of rate of lime addition. As this system approached its optimum, the size of the incremental changes in lime rate were reduced to prevent excessive overshoot. In addition, a small reversed polarity signal was occasionally introduced to confirm that the process was actually approaching its optimum.

I delivered two lectures to the Academy of Sciences. They were entitled: 1) "Adaptive, Optimizing, and Learning Systems," and 2) "Design of Systems with Restrictions." My lectures were well received by audiences of 400 and 200 and there were many questions afterwards.

I talked with Professor A. N. Miljakh, Director of the Electrotechnical Institute of the Ukrainian Academy of Sciences, USSR, Chkalov St. 55B, Kiev; he had a Ph.D. degree in Technical Sciences. He would like to see an exchange of graduate students between technical institutes in Russia and their equivalent in the United States. In particular, he would be willing to consider discussions concerning the exchange of graduate students between his institute and the University of California. He does not believe that men at the professional rank should be exchanged.

Professor Ivakhnenko had arranged for me to see the Rector of the University of Kiev, but we were never able to get together. I visited an analog computing facility in the University, which consisted of a mechanical differential analyzer of approximately 24 large mechanical integrators and 4 mechanical multipliers, all interlinked with electrical-synchro systems. They claimed 5-digit accuracy for each component, and 4-digit accuracy for the results of complex problems. It did not appear that it was being used at present.

In Kiev is the Institute of Automation of the Ukraine under the direction of Professor P. M. Melnik, which I did not have an opportunity to visit. I understand that in this institute there are approximately seven hundred people, of whom one hundred and forty are scientists, and two hundred and eighty are engineers and technicians. They have a Russian M10 analog computer, and are working on the automation of power systems, the automation of gas and steam turbines, of nonferrous metal production, programmed control of machine tools, telemetering, chemical analysis, blast furnaces, open-hearth furnaces, special gauges and transducers, and a time-shared multipurpose sampled-data system.

Kharkov

I intended to call on Professor Vashura, who is in automation at the Polytechnic Institute in Kharkov. When I arrived he was not in, but I did talk with Docent Paul Stupel and Engineer Constantin Didenco. They have ten thousand students in this Institute, of whom three thousand are electrotechnic students, and there are six hundred electrotechnic professors and docents. They have work in machine design, apparatus design, automatics and high voltage. They had two analog computers. One had twenty-four amplifiers that looked something like an early model Philbrick computer, and the other had sixteen amplifiers and was essentially of portable design. They also had a digitally controlled, two-dimensional contour milling machine. They had an elaborate pneumatic control board with examples of all the available pneumatic controllers and recorders manufactured in Russia, and provision for measuring the characteristics of these controllers and of connecting them to models of processes. They had no means for generating low-frequency sine waves. Their theoretical analyses were usually based on step-response curves and not upon frequency response.

They had typical small positional servos in their laboratory. Considerable emphasis was being given to the construction of ac and dc control amplifiers using transistors. The total space available for automatic control was about 800 square meters, excluding the machinery laboratory. This Polytechnic Institute looked more like a German or western technische hochschule than any other that I saw in Russia. The different departments had separate buildings of a few stories in height, and the whole occupied several city blocks.

In the Soviet Union, the University of Kharkov is second in size only to the Moscow State University. A new building, copied after the skyscraper design of the Moscow University, but smaller, was being completed. Here they had departments of mathematics, chemistry, geology, and physics.

I visited the equivalent of a technical high school in Kharkov. The school had a three-year curriculum for students who were in their twenties. These students had been working in industry from the age of fourteen. They came to school to learn a specific trade. There were departments of chemistry, physics, mathematics, electricity, industrial design, and testing, that I was told about. The test laboratory had machines for testing of metals in torsion, tension, compression, impact, hardness, and photomicrographic inspection, all of very fine quality. In an adjoining laboratory, equivalent machines were available for testing concrete. The electrical laboratory had not yet been built, and students were using a laboratory in another institution. There seemed to be one professor in charge of each laboratory. The industrial design laboratory seemed to be concerned with building small models of factory layouts, process layouts, and machine installations.

We had difficulty driving through Kharkov because the highway through the industrial part of the city was closed to foreigners. We mailed many letters here. Those to the U. S. arrived, but those to people in Russia had not arrived several weeks later.

Travel Observations

There is a continuous microwave channel from Yalta to Simferopol, Melitopol, Zaporozhe, Kharkov, Kursk, Orjel, Moscow, Kalinin, Novgorod, and Leningrad. In the south of Russia, there was a second micro-

wave system being built parallel to the first. There seemed to be electric power available everywhere, and there was little evidence of a power shortage in the form of flickering lights, intermittent service or low voltage. In Simferopol, there was a power failure of about one hour after a rainy morning. The restaurant had only cold cuts because the kitchen was electric. The garage grease guns would not work without 3-phase power. I would guess that the average country house had at least one light bulb and one radio. The apartments of the professors in the cities had TV, phonograph, tape recorder, and radio, but no electrical appliances. TV reception in Kiev was of good quality, superior to that which I have seen in Berkeley. The very best homes had small two- or three-cubic foot electric refrigerators. They had two-burner kerosene or propane stoves.

Some of the camp grounds had diesel electric generators rated 25 kilowatts delivering 3-phase power, which was 220 volts from line to natural. These small stationary power plants had manual control of speed and manual control of voltage. They had no automatic regulators.

Near Melitopol, there were oil fields and a military installation. In Zaporozhe we visited the first hydro-electric dam and plant built after the communist revolution. In this town there were many steel mills, and at night the sky was lighted as it is in Pittsburgh, Pennsylvania, by the pouring of steel. Kharkov was a great rail center, and we heard the trains rolling all night long. In Orjel we were next to a military jet airfield and the jets flew, and were being tested, all night long. The agricultural crop of the Ukraine was primarily wheat, and in the large cities there were big wheat elevators. We also passed concrete plants in several places. In Belgorod, north of Kharkov, we passed road workers digging a ditch under the supervision of armed guards. There were police at every major road intersection in the country, who checked the papers and passports of many of the Russian trucks. There were police at almost every corner in the cities, and several at each major intersection. They stood like ornaments, except that they decided who was wrong at each accident.

On the road between Yalta and Moscow, there were only a few stations where one could have his car lubricated. In the camp grounds, there were racks available for an individual to lubricate his own car. There seemed to be little attention given to the maintenance of the state-owned trucks which were the major traffic on the highway. Two-thirds of the trucks going in either direction were empty except for the fifty-gallon barrel of spare gasoline which each truck carried in back. Gas stations were about 150 miles apart. Many of the trucks sounded as though the differentials were out of oil. We saw almost no farm machinery, and the few pieces which we did see were badly in need of maintenance and painting.

The recreational facilities which we saw, both available and extensively used, were the beaches and waterways. There was much swimming and there were quite a number of boats. In Kiev there were approximately 2000 boats, many of which had 5-horsepower inboard engines of a standard Russian design. All of these boats are built personally or by private enterprise, and not by a state factory. The boat owners in general did not own a bicycle or an auto.

All books, magazines, newspapers, radio, and TV were completely controlled by the government. The Voice of America broadcasts are apparently the one sinful source of information for the Russians. In several cities the reception was good, and no serious attempt was made to jam the signals.

Moscow

At the opening session of the Congress in Moscow we were provided with small transistorized FM receivers and headsets for simultaneous translation of the opening ceremonies. There were perhaps 800 of these receivers available with four channels which could be switched at the desire of the listener. These were used only for the plenary sessions. The technical sessions were given in Russian or English with a translator who would translate sentence by sentence. The translators were specialized and they were chosen for each paper on the basis of their specialty.

The Russians had received the Congress papers nine months in advance and had prepared Russian translations. They also had assigned a reviewer for each paper and had prepared a critical discussion for each paper. The Russian papers were not made available to others until one week prior to the Congress, and there was little prepared discussion of them.

At the Congress I spoke on "Philosophy of Control." At the Polytechnical Museum, I gave five lectures entitled:

1) "A Magnetic Delay Line for a Dead-Time Analog,"
2) "The Control of Flux Phase in an AC Machine,"
3) "Maximum-Effort Controls for Second-Order and Higher-Order Systems,"
4) "Testing with a Complex-Zero Signal Generator,"
5) "Root Locus Techniques and Complex Function Computer."

In Moscow I visited the Institute of Automatics and Telemechanics of the Academy of Sciences under the direction of Professor V. A. Trapeznikov. This Institute has a research staff of approximately five academicians, 25 doctors, which are honorary degrees, 100 candidates, equivalent to our doctor's degree, and 270 diploma engineers, making a total of 400 technically trained personnel. In addition, there were 400 assistants and secretaries. We were told that they had a budget of 1,000,000 rubles per year, but other groups were told that they had a budget of 100,000,000 rubles per year. Some of their staff who are well known are: A. A. Feldbaum, adaptive and extremum controllers, Y. Z. Tsypkin, sampled- and quantized-data systems, V. S. Pugachev, random signals, M. A. Aizerman, pneumatic and nonlinear elements, A. M. Letov, mathematics, and Professor Lehner, self-adaptive systems. We saw some analog computers with approximately 18 amplifiers in each of two, and a third with 12 amplifiers. One of these computers used a dc chopper amplifier in parallel with an ac amplifier. We saw a two-variable adaptive system where the variables were perturbed by step changes and the direction of optimum change was calculated after each pair of perturbations. We saw a 12-variable adaptive system whereby each of the 12 variables was perturbed by a step change in sequence, and then after an appropriate computational period, all the variables were perturbed in synchronism by steps of sizes so chosen as to cause the entire system to approach its optimum adjustment by a route of steepest ascent. We saw research on a dead-tine control for a four-stand steel rolling mill, which used a magnetic tape unit to represent the dead time in the steel mill. The control was based on a feed-forward predictor calculating the best adjustments of the rolls on the basis of an X-ray gauge following the first roll, and an additional feedback system, which made small corrections in the computation parameters based on measurements from an X-ray gauge following the last roll. We saw conventional force-balance pneumatic controllers and pneumatic multipliers and function generators whose principle was not explained to us. We saw a pneumatic bistable relay using a miniature airfoil in a jet.

The laboratory rooms were very small and each room had little equipment except the piece of apparatus being used in the experiment. We saw no low-frequency sine-wave generators. We saw no evidence that root locus techniques were used in their analyses or syntheses. The laboratories did not look as though they were actually worked in, but looked more like rooms in which display equipment was assembled for the purpose of persuading prime contractors to consider these ideas in their plants or projects. We were told that the Institute carried the design work all the way from invention to the final design of the piece of apparatus to be installed in the factory. We saw no drafting facilities nor model machine shop adequate to perform these functions.

The Institute of Automatics and Telemechanics publishes a magazine entitled *Automatics and Telemechanics*, which distributes approximately 6000 issues inside of Russia, and approximately 2000 issues outside of Russia. They say that approximately 1000 engineers each year come to them for consultation on control problems. Their staff has no chemists. They do not have a digital computer, but the central computing facility in Moscow is available to them when needed with a waiting time of perhaps one day. They cooperate with the Institute of Complex Automation and Mechanization, whose purpose is to develop complete systems of control for industry. This latter institute is under the direction of E. P. Stefani. We inquired about the exchange of personnel between institutes and were told that leave was granted only when "necessary." Under more specific questioning, they were unable to give any example of a case where an engineer from one institute had gone to another institute for research or for an extended period of consultation. We gathered that even short time visits were discouraged. Special permission has to be obtained for an engineer living in one city to go visit another city.

Education

I visited the Moscow Power Institute, which is the largest polytechnical school in Russia. It has an enrollment of approximately 15,000, with approximately 3000 in automatic control. Mr. Bondin, an exchange student with the University of California a few years ago, was my guide at the Moscow Power Institute. Their control laboratories seemed to have slightly more equipment than our teaching laboratories at the University of California, which has one-twentieth the number of students. In the control field, they had ac synchro-systems, servo boards for dc positional controls, pneumatic controllers, automatically balancing potentiometers, and machine controls. We did not see any work in root locus methods, maximum effort or relay controllers, hydraulic systems, or analog computers. They did not use Bode plots or frequency-response methods. They used step-response tests in the laboratory.

This institute is like a cooperative school in that the students spend half of their time in industry and require a minimum of 5½ years for a diploma. The living allowance for students is 300 to 500 rubles per month, depending on grades. The diploma engineer is assigned his first job at which he must work for 3 years.

The graduate school has only about 250 students, who are paid 800 to 1000 rubles per month by the employer from whom they have a leave of absence. After three years, they receive the title of "Candidate of Technical Science." This is supposed to be equivalent to our Ph.D. degree, but I was quite unimpressed by their research, and I would say that the Candidate is equivalent to our M.S. degree. After five years of graduate work, a student receives the title of "Doctor of Technical Science." Fifteen per cent of the students do not finish their candidate degree.

The government can staff its new cities and factories in the interior by assigning new graduates to their first jobs. After three years, the worker is permitted to move, but it is difficult for him to find work except in the new industries, and he is therefore virtually frozen in his job.

Typical income per month figures were: minimum wage, 350 rubles, secretary, 700 rubles, cutter in shoe factory, 800 rubles minimum—2000 rubles maximum, taxi driver, 1000 rubles, secondary school teacher, 1100 rubles, machinist, 1200 rubles, theater business manager, 1500 rubles, 42-year-old mechanic, 1800 rubles, and skilled fitter, 2000 rubles.

Typical prices were:

6 r.	gallon gasoline
1.3 r.	loaf of bread
7.5 r.	can of milk
2.2 r.	ice cream bar
3 r.	factory lunch
8–20 r.	restaurant dinner
35 r.	pound of coffee
20 r.	pound of cheese
15 r.	pound of butter
12 r.	pound of cherries
40 r.	pound of chocolate
30 r.	pound of caviar
100 r.	yard of cheap cotton
80 r.	dime store lady's straw hat
180 r.	used German drawing instruments
400 r.	used stylish hat
400–800 r.	pair shoes (men or women)
1100 r.	used English horn
2000 r.	used Contax camera
2000 r.	man's cheap suit
2500 r.	small TV
150,000 r.	small automobile
600 r.	State resort, per person per month
10,000 r.	table, chairs, buffet, 4 beds, radio, TV
4 r.	camp ground admission per person per day
100 r.	apartment rent for salary of 1500 r./month (and utilities)
240 r.	apartment rent and utilities for salary of 4800 r./month
900 r.	very frugal monthly grocery bill for 2 parents and 2 children
2400 r.	monthly grocery bill for 4 adult workers living together
40 r.	black market exchange rate per dollar.

There was a universal income tax of 10 per cent and, in addition, a bachelor's tax and a small family tax. It can be seen that the purchasing power of a ruble in Russia compared with a dollar in the U. S. for the same quality merchandise varied between 10 and 50 rubles per dollar for Russian goods, and 70 rubles per dollar for automobiles and imports. For services, rent and utilities, books and public transportation, five to ten rubles were equivalent to a dollar. For example, a taxidriver and his wife together could earn 1800 rubles. Rent, taxes, and food would take 1170 rubles, leaving 630 rubles for household, clothes, and all other purchases per month. A full professor earning 6000 rubles would pay 400 rubles for taxes, rent, and utilities. An adequate diet for a family of 3 would cost at least 3000 rubles. This leaves 2100 rubles for household, merchandise, and all other purchases each month.

There were 1000 on the entire teaching staff of the Moscow Power Institute. This is a student-faculty ratio of 15. A typical professor received 6000 rubles per month, of which 4000 came from academic funds and 2000 came from research grants from industry. The very best professors and department heads received 10,000 rubles per month. For comparison, the income of graduate librarians, waitresses, and Intourist translators was 700 rubles per month. We met several Intourist translators and administrators who were also school teachers in order to receive two salaries.

The children have an active political life. We were told that at the age of 52 days, 95 per cent of the young babies in all Russia are placed in state day nurseries so that the mother can go back to work. In the nursery schools they teach them to follow the leader, to march in line, and to participate in group activities. Stories are read to them. There was no daydreaming or imaginative play. We visited a nursery school in which the toys were kept on display for the parents in the lobby, but the children did not play with them. The teachers were dedicated and patient. The three-year-old children join the Octoberites. Most of the seven- to fifteen-year-olds belong to the Young Pioneers. The State reserves for itself the right to give all instruction. The children are taken on many excursions in groups to museums of atheism, communist industrial exhibitions, agricultural exhibitions, art museums, propaganda exhibits, halls of culture, historical landmarks, swimming, and marches in the country. The children do not have or use play equipment. They do not have building blocks, dolls, small cars or planes, sand piles, swings, teeter-totters, roller skates, wagons, tricycles, or bicycles. Although dolls are available in the stores, we never saw a girl with a doll. They have paper dolls and charming children's books with activities and illustrated stories. We never saw children making roads in the dirt, or pretending that a block of wood was a wagon or an airplane.

In the country we saw grown people going on picnics and excursions in trucks, not as families. On Sunday in Moscow, we saw family groups walking together. The parents are sweet and affectionate with their children. Young girls are particularly protected.

The children also go to school. In the elementary schools, first grade does not start until the pupil is seven years old. The teachers are not unduly demanding for the first five years. Then, between the ages of 12–14, the competition becomes very severe. The student's grades during this time determine whether he will have any further education. During this time, the student applicants are accepted into special schools, such as an all-English school in which all classes are conducted in English. Many students go to work at the age of 14 for a year or two, even though they have been accepted for further schooling. Before admission to a university, a student must have a year or two of practical experience. Some make an attempt to fit this experience to their future field of interest; thus, a potential chemist planned to spend a year washing test tubes after high school graduation, and an Intourist guide was planning to attend a language school.

Correspondence schools were available to those not admitted to high school or college. We met several who had studied English this way.

There were pedagogical schools for training teachers in some of the smaller towns. The English student teachers found Intourist a good place to work to gain experience.

I visited the experimental physiological laboratory of the Institute of Psychiatry of the Academy of Medical Sciences, which was located on the grounds of a mental hospital in Moscow. The director was Professor S. N. Braines. They were performing learning experiments with mice, rats, dogs, monkeys, and a chimpanzee. Normal maze learning graphs of errors versus time were shown to have three regions, the first two of equal length in time: my interpretation of these regions were 1) a learning region in which the graphed curve decreased asymptotically to a low value of errors, 2) a confirming region in which the errors stayed constant at this low value, and 3) an automatic response region in which the errors were significantly less than in the confirming region. I hypothesized that in the final region the learning mechanism was turned off in order to achieve faster and more accurate response. Experiments were performed of learning rate under the influence of various drugs and human serum from mental patients. Experiments in blocking and unblocking conditioned reflexes were performed on monkeys. They were starting work on an electrical model for simple conditioned reflexes.

Driving toward Leningrad, we paralleled their 18-inch gas line. In Novgorod, even 612 didn't work. We saw churches that had been converted into apartments. They had

a set of bells in their kremlin that had never been hung because they were too heavy for their bell tower. TV antennas were numerous in the areas around the cities. The number of houses with antennas was similar to the U.S.A., but the number of antennas per house was five times as many.

Leningrad

We stayed at the campground in Sestroretsk, but I also got a hotel room in town for headquarters. A man offered to buy our car, our clothes, or anything that we would sell. Money changers were selling rubles at 40 per dollar, but we always refused because this was illegal. (The recent currency change will wipe out the fortunes of Russians who cannot explain where they got their money.)

In Leningrad I visited the Electromechanical Institute of the Academy of Sciences administered by the Director, Academician M. P. Kostenko, and Deputy Director, Dr. A. A. Voronov. This is a research institute with a total staff of 350, of whom 13 are doctors and 30 are students for the doctor's degree. This Institute works cooperatively with another research institute in Novosibirsk which has departments of automatic and metallurgical sciences and power systems and traction.

The pedagogical institutes in Leningrad are the Leningrad Electrotechnical Institute, with 8000 students, and the Leningrad Polytechnical Institute, which is 15 kilometers from the city, and has a student body of 9000. The Leningrad University has physical sciences, but not engineering. I met two men walking by the University. One was an electrotechnician, 25 years old, working in guidance with a salary of 1100 rubles per month; and the other was a physicist, 32 years old, working in photographic optics at a salary of 2000 rubles per month. Both of these men were in an unnamed research institute. I also met a Professor-Doctor Usman Effendi, Professor of Indonesian language and literature, who had been at Leningrad State University for several years, and who returned to Indonesia only for his holidays.

The Academy of Sciences has departments of: 1) technical science, including engineering, 2) biology, 3) astronomy, 4) applied physics and mathematics, 5) chemistry, 6) geology and geography, 7) physiology, 8) history, 9) economics, 10) philosophy and law, 11) literature and language. The Electromechanical Institute in Leningrad is one of the institutes under the Academy of Sciences. A typical engineer here was Miss A. Dernvo. She graduated from a five-year program in one of the pedagogical institutes whose curriculum was approximately as follows:

First year: mathematics
beginning physics
chemistry
Russian language
literature
English

Second year: differential equations
complex algebra
general physics
theoretical mechanics
electrotechniques
construction
dialectic materialism
English or German

Third year: dc machines
theory of ac circuits
thermodynamics and machines
hydraulics
dialectic materialism
English or German

Fourth year: hydraulic machines
ac machines
electrical nets
electrical stations

Fifth year: electric traction machines
distribution systems
ignitrons
radio techniques
advanced electric stations
construction of electrical machines.

In the last half of her fifth year, Miss Dernvo engaged in diploma work, which is individual research. Her project was collector machines, and the title of her paper was "Compensation Winding on Stator for Alternating Current Commutator Machine with High Power Factor." She was continuing to work on commutator machines.

In addition to the salaries of 1000–2000 rubles per month of engineers in these research institutes, it is possible for an exceptional engineer to win a bonus. A premium in the name of Lenin carries a stipend of 100,000 rubles for an exceptional contribution to the Soviet Union. There are, in addition, small premiums ranging from 500 rubles up for outstanding contributions.

I delivered two lectures entitled "The Control of Flux Phase in an AC Machine," and "Nonminimum Phase Feedback Control Systems."

In addition to visiting Professor Voronov, the Director of the Electromechanical Institute, our entire family was invited to the dachas of Professor Alekseev Aleksandr, Academician Michael Kostenko, and Academician Zavalishina. Each of them hospitably showed us their acre in the pines and their gardens.

The Electromechanical Institute has completed the translation of my book "Feedback Control Systems" into Russian under the direction of Professor E. P. Popov.

One activity of this Institute is the supervision of the electrification of the railway system. In the U.S.S.R., a total of 9000 km are now electrified. There are 4300 km electrified at 3000 v dc, and 1200 km electrified at 25,000 v, 50 cps. They have 50 locomotives of French manufacture, and they have ordered 25 locomotives of German manufacture, using semiconductors for converting the 50 cps to dc for the traction motors. Their plans for all future expansion are to use high-voltage ac distribution with semiconductors in the locomotive.

This institute is also making the plans for long-distance transmission of power from Dondas to Stalingrad, and from thence to Moscow. The distance from Dondas to Stalingrad is 500 km, and this will be an 800-kv dc transmission at 900 a, 750 Mw. One power line will be 400 kv, positive to ground, and the other power line will be 400 kv, negative to ground, with constant-current drive so that under fault conditions the current is limited and the voltage drops to zero. Under variable load the current is kept constant at 900 a, and the voltage is raised or lowered. Constant current is maintained on the line by an extinction angle regulator of the inverters in Stalingrad.

The Stalingrad to Moscow line will be approximately 1000 km in length, will operate at 500 kv, 50 cps, and will carry approximately 1800 Mw. The stability of this long ac line will be enhanced by an unusual control of the power input from the dc transmission system. For increasing torque angle on the ac line, the power input from the dc line will be correspondingly diminished. This will permit operation at torque angles greatly in excess of the usual stability limit of 90°. To test this proposed system, a complete model of both transmission systems including the rotating machinery at Donads, Stalingrad, and Moscow has been built in the Institute to the scale of 2 kv = 1 per unit voltage, 10 a = 1 per unit current, and with the moments of inertia of the model machines so chosen that their time constants are identical to the time constants of the large equipment. This permits the use of the actual voltage regulators and other auxiliary equipment in the model, so that the dynamic characteristics of the model are identical to those of the system proposed.

The high-voltage power lines will use three-conductor bundled conductors with 30-mm wire diameters and 40-cm spacing of the conductors in an equilateral triangle. The phases will be separated by 10.5 meters with 5.5 meters minimum clearance to the tower. The line will have 8 meters minimum distance above ground, and maximum span length of 450 meters.

The professors at the Electromechanical Institute of Leningrad were very much interested in the possibility of exchange professorships in the United States. They felt that a contract similar to the one negotiated between the University of California and the University of Moscow should be negotiated to provide for the exchange of engineering professors. They suggested that the appropriate contracting agency would be the Department of Polytechnical and Electrotechnical Schools in the Ministry of Higher Education in Moscow. If such a contract were negotiated, the professors would be interested in discussing the implementation of it through specific proposed exchanges with the University of California.

I visited the Vibrator manufacturing plant in Leningrad. Vladimir Ramonovsky, Chief of the Construction Bureau, was my host. They manufacture high-precision laboratory dc instruments of 0.1 per cent accuracy, galvanometers, ac and dc switchboard instruments, high-frequency thermocouple instruments for use up to 200 Mc, mechanical multichannel oscillographs, photoelectric amplifiers not using servos, photocells and photo resistors, electrostatic meters, camera exposure meters and small

components such as springs and suspensions for use in other industries. Their staff consisted of 550 technically trained men and engineers, 250 administrative personnel, clerks, and secretaries, and 2200 factory workers and production line employees, making a total personnel of 3000. Their production had a value of 200,000,000 rubles per year, which was approximately 66,000 rubles per year per man in value. The average salary in the factory was 980 rubles per month, which is 11,760 rubles per year. The production per man is therefore six times the salary per man.

They produce 130 different types of instruments with an average sale price of 660 rubles per piece. They produce a total of 300,000 pieces per year.

Each worker has a base salary, plus a piecework salary, plus bonuses for very exceptional production. The best workers can make as much as 1500 or 1600 rubles per month. Each worker works seven hours a day for five days, and six hours a day on Saturday for a total of 41 hours per week. (The seven-year plan provides that in 1964 the work week will be reduced to 36 hours.) They receive a two-week minimum vacation and skilled technicians receive three- and four-week vacations. The in-plant training varies from several weeks to several years before the worker is put on the production line, according to the production manager. He also said the minimum age was 18 years, but some of the girls on the production line of small instruments appeared to be about 14 years old. Twice a day, they have calisthenics in the entire plant. They have won a prize for interplant competition in calisthenics. They have a staff of consultants, including one academician. They have on their staff two engineers who are Candidates, and the Director is also a Candidate. Ninety per cent of their engineers are Diploma engineers, and about 300 of them were trained in Leningrad. The rate of increase of plant production has been a doubling of output each three years.

They said they had a program of research on new instruments, and that they produced only their own designs. I did not see any of the research laboratories. They were making a line of instruments which depended upon a bronze ribbon torsion suspension rather than pivots and jewels. They were making galvanometers, light-beam instruments, low-precision laboratory instruments, and high-precision standards all using these suspensions. They had invented their own impact ball die machine for fabricating their bronze ribbons. They made a photoelectric photometer which used a selenium cell 4 cm by 4 cm. They claimed 10 per cent rejects on their selenium cells. The photo-exposure meter used a winding of 3000 turns of 0.02-mm wire, external steel pivots, and ruby bearings. They had invented their own method for grinding these pivots, and for induction heating and tempering them with a 10-Mc induction heater. The exposure meter had a complex mechanical cam to convert the meter reading into equivalent light illumination values. This meter sold for 650 rubles. Their 5-×5-inch plastic meters from 1 a to 10 a sold for 600 rubles. Their laboratory precision instruments sold for 1000 rubles, and their light-beam instrument with 1 per cent accuracy and 0.1 μa full-scale sold for 800 rubles.

I saw the manufacturing of the bronze suspension ribbons, the manufacturing of meter springs, the grinding of pivots, and the assembly of small meters. All of these used young girls dressed in white dresses and white caps on the production lines. The rooms were clean, and well lighted. The girls' manual dexterity appeared to be average. I asked who was responsible for correcting meters that did not meet final inspection, and was told that this never occurred. The meters are arranged with a magnetic shunt for fine adjustment of the calibration, and demagnetizing of the permanent magnet was used for gross adjustments of the calibration after the meter was completely assembled. They had a good automatic machine for inscribing a meter dial which would interpolate between a series of calibration points. All of the lettering on their dials was glued-on plastic cut-out letters rather than ink.

In the Director's office at the Vibrator plant, there were seven telephones on his desk. These were of various sizes and shapes. I saw no telephone switchboard or PBX board. As was customary with each telephone in Russia, the telephone number was not shown on the instrument. In addition, no telephone book was available.

In Leningrad I visited a basement TV and radio repair shop. The proprietor had tape recorders manufactured in Kiev that were similar to German designs that I had seen. His was a private enterprise, selling both new and used TV's and radios and the repair of the same. He had small quantities of war surplus material for sale to hobbyists. This was one of the few private enterprises that we saw in all Russia. The manager showed me a large wound that he received in the last war and said that he would never again fight in a war.

When we left Leningrad, we drove north toward Vyborg. All highway traffic in both directions was stopped at a permanent roadblock south of Vyborg, where our documents were examined carefully. In the Vyborg railway station, we changed our remaining rubles back to dollars, paying a one dollar fee. Leaving Vyborg, we were again stopped at a permanent roadblock. Beyond this point there were only a few habitations. Exit customs was not difficult. After being cleared, we had a personal 2-man motorcycle escort which took us precisely to the border line and made certain that we entered Finland promptly.

The next month was spent visiting the Scandinavian engineering institutions. The wealth of buildings and goods in Helsinki made us acutely aware of what we had not seen in Russia. The abundant books, magazines and newspapers in all languages were especially welcome. The serene, happy and confident countenances of the people reflected the rights and responsibilities of each individual.

Otto J. M. Smith
University of California
Berkeley, Calif.

Inverse Root-Locus, Reversed Root-Locus or Complementary Root-Locus?*

The root-locus method has found wide application during the last decade in the analysis and synthesis of linear lumped stationary feedback systems. The author feels that there is a need for a standard and consistent terminology while referring to the locus of the open-loop poles when the overall system is specified.

The root-locus method, as suggested by Evans,[1] is for the determination of the closed-loop poles of a unity negative feedback system from the location of the poles and zeros of the open-loop system (referred to, here, as the analysis problem). This has become an integral part of most textbooks in the field published today, wherein different root-locus shapes and compensation methods using the root-locus approach are discussed for positive values of the gain parameter K.

When K takes on negative values the root-locus applies to the positive feedback case, and Yeh[2] has discussed various root-locus shapes for $-\infty < K < \infty$, calling the root-locus for negative values of K the 0° locus and that for positive values of K the 180° locus.

In 1956, Aseltine[3] observed the symmetry between the analysis and synthesis problems of positive and negative feedback systems and suggested the "inverse root-locus method." This involves the determination of the open-loop poles from a knowledge of the closed-loop singularities by drawing what would amount to the conventional root-locus for a positive feedback system. The symmetry between the analysis and synthesis problems (Fig. 1) was independently observed by the author[4] (at that time working on his doctoral thesis at Harvard University), who used the "complementary root-locus" for the synthesis problem. At about the same time Zaborsky,[5]

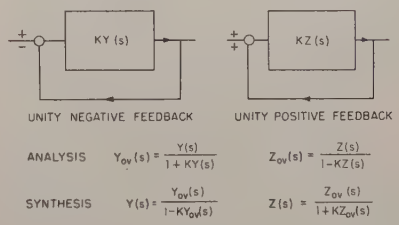


Fig. 1.

* Received by the PGAC, December 5, 1960.

[1] W. R. Evans, "Control system synthesis by root-locus method," *Trans. AIEE (Commun. and Electronics)*, vol. 69, pp. 66–69; 1950.

[2] V. C. M. Yeh, "The study of transients in linear feedback systems by conformal mapping and the root-locus method," *Trans. ASME*, vol. 76, pp. 349–361; 1954.

[3] J. A. Aseltine, "Feedback system synthesis by the inverse root-locus method," 1956 IRE National Convention Record, pt. 2, pp. 13–17.

[4] K. S. Narendra, "Synthesis of Linear Feedback systems through Pole-Zero Configurations" Doctoral Dissertation, Harvard University, Cambridge, Mass., January, 1959.

[5] J. Zaborsky "Integrated s-plane synthesis using two-way root-locus," *Trans. AIEE*, vol. 75 (*Commun. and Electronics*), pp. 797–801; January, 1957.

working independently on the same topic, referred to the locus as the "reversed root-locus."

Though the work by all three was carried out independently, Aseltine's paper was the first to appear and subsequently his terminology has been used. The author feels that this root-locus will find extensive use in the future, along with the conventional root-locus, in the design of feedback systems. Since this method is relatively new, a standard and consistent name should be given to it for the sake of uniformity in all future publications.

The title "inverse root-locus" might imply that it is the root-locus of the inverse transfer function, which it is not. The term "reversed root-locus" is certainly more descriptive, though not justifiable analytically. The terms 0° locus and 180° locus, suggested by Yeh, are certainly satisfactory, but the latter is already well known as the "root locus." The term "complementary root-locus" is proposed by the author since the conventional root-locus and the complementary root-locus together form a complete algebraic curve.

K. S. Narendra
Harvard University
Cambridge, Mass.

New Method of Compensating Network Design for Feedback Systems*

In the design of a feedback control system, it is often necessary to employ a compensating network in order to attain the desired stability margin. To determine a proper series compensating network, either the Nyquist plot or the Bode plot can be used. Of the two, the Bode plot is preferred to the Nyquist plot mainly because with the latter, the design process involves a larger amount of trial and error work. However, the Nyquist plot gives a clearer physical picture and a better insight into the system performance than the Bode plot. The new design method described here uses the Nyquist plot as its basis, but cuts the amount of design effort appreciably.

The basic principle of the new approach is explained using the simple feedback control system of Fig. 1. This is done for the sake of clarity, but in no way means limited application to more complicated systems. In Fig. 1, $G_s(s)$ and $G_c(s)$ are the transfer functions of the system and the compensating network, respectively. The Nyquist plot of the same system is shown in Fig. 2. The forward transfer function $G_s(j\omega)$ plot shows that the feedback control system is unstable unless a proper compensating network is used. With the compensating network, the resultant $G_c(j\omega)G_s(j\omega)$ plot has a positive phase margin indicating that the feedback system is now stable.[1]

Suppose that it is required that the resultant plot has a phase margin of γ degrees, or in other words, that the resultant plot passes through point G_0 on the unit circle as shown in Fig. 2. Here, G_0 is a complex number. Then the problem becomes that of designing a compensating network satisfying this requirement.

Suppose also that the compensating network has a Nyquist plot shown in Fig. 3(a). The inverse of the compensating network, $1/G_c(j\omega)$, has a Nyquist plot shown in Fig. 3(b). Using this, it is very easy to construct the $G_0/G_c(j\omega)$ plot since it simply means a rotation of the $1/G_c(j\omega)$ plot by $(180+\gamma)$ degrees as shown in Fig. 3(c). The $G_0/G_c(j\omega)$ plot has a characteristic that any point on this plot can be brought to G_0 through the use of the compensating network $G_c(j\omega)$. This is obvious but important because it is the very basis of the new approach.

Fig. 3(c) shows the $G_s(j\omega)$ plot together with the $G_0/G_c(j\omega)$ plot. The two plots have two intersecting points a and a'. Either of the two points can be brought to G_0 by the compensating network. If this is done, the resultant $G_s(j\omega)G_c(j\omega)$ plot passes through point G_0, meaning that the compensating network satisfies the requirement. Choosing point a, this condition is realized when the $G_s(j\omega)$ plot and the $G_0/G_c(j\omega)$ plot intersect at a at the same frequency. Since $G_s(j\omega)$ is given, point a corresponds to a certain frequency. Then the procedure is to adjust $G_c(j\omega)$ so that point a also means the same frequency for this function. This is all that is necessary to design a compensating network satisfying the phase margin requirement. Point a' can be used as well instead of point a. The number of the intersecting points of the $G_s(j\omega)$ and the $G_0/G_c(j\omega)$ plots is not necessarily two in all cases. It can be more than or less than two. If there is no intersecting point, however, it simply means that the compensating network cannot satisfy the phase margin requirement.

Although, in the preceding illustration, point G_0 with phase margin of γ degrees was chosen for the resultant plot to go through, any arbitrary point can be selected and the same procedure used to design a series compensating network. It is also evident that a similar approach can be made using the Nyquist plot of the inverse transfer function, $1/G_s(j\omega)$. In this case, the $1/G_s(j\omega)$ plot and the $G_c(j\omega)$ plot are made and intersections observed.

Two simple examples are given for clarification of the new method.

A. Phase-Lag Compensating Network

Suppose that $G_s(j\omega)$ of Fig. 4(c) is the Nyquist plot of the system and it is required that the phase margin of γ degrees be obtained using a phase-lag compensating network. The transfer function of a phase-lag network is given by

$$G_c(s) = \frac{1 + sT_2}{1 + sT_1} \qquad T_1 > T_2.$$

The ratio of T_2/T_1 and either one of T_2 and T_1 specify the function completely. In Fig.

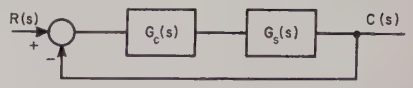


Fig. 1—Feedback control system with a series compensating network.

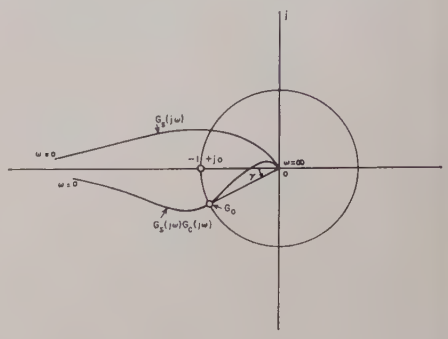
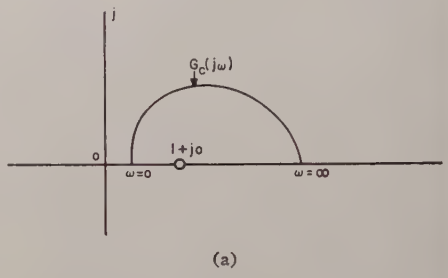
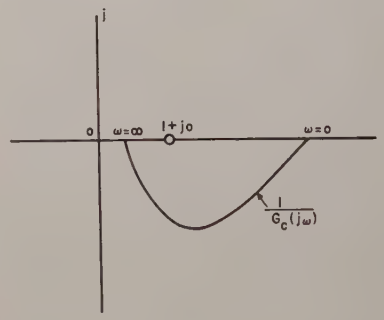


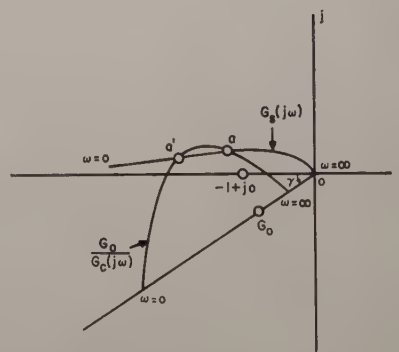
Fig. 2—Nyquist plots of the system function and of the resultant function containing the compensating network.



(a)



(b)



(c)

Fig. 3—(a) $G_c(j\omega)$ plot. (b) $1/G_c(j\omega)$ plot. (c) $G_0/G_c(j\omega)$ plot and $G_s(j\omega)$ plot.

* Received by the PGAC, August 9, 1960; revised manuscript received, March 7, 1961.

$_s(s)$ is assumed to be stable by itself.

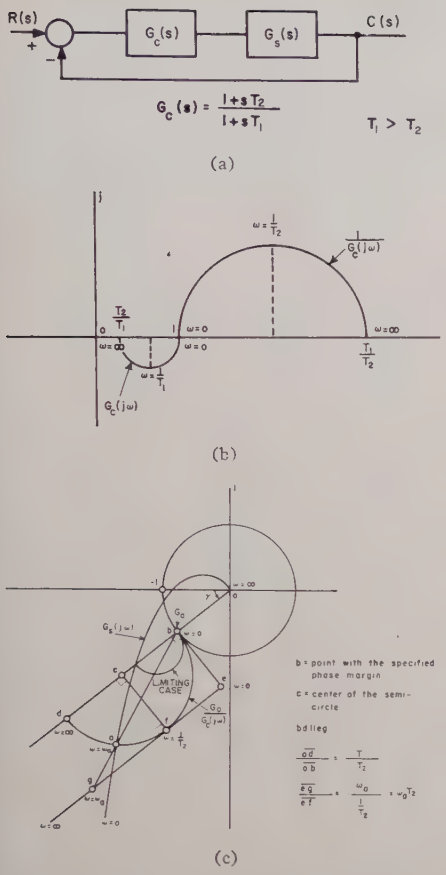


Fig. 4—Design of a phase-lag series compensating network.

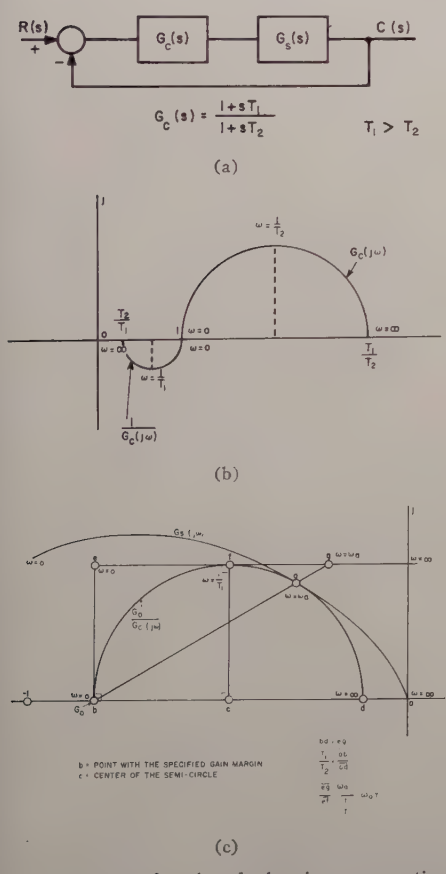


Fig. 5—Design of a phase-lead series compensating network. Fig. 5(c) is drawn on an expanded scale to show the details.

4(b), the $G_c(j\omega)$ plot and the $1/G_c(j\omega)$ plot are shown. These plots are semicircles with points $\omega = 1/T_1$ and $\omega = 1/T_2$ conveniently located. The $G_0/G_c(j\omega)$ plot bfd of Fig. 4(c) is very easily made because it is also a semicircle.

Now the ratio of the two time constants is given by

$$\frac{T_1}{T_2} = \frac{\overline{od}}{\overline{ob}} = \overline{od}, \qquad (\overline{ob} = 1).$$

The semicircle must be drawn so that it intersects the $G_s(j\omega)$ plot, as otherwise no practical design is possible. This means in this case that there is a permissible lower limit for T_1/T_2 but no theoretical upper limit. The $G_s(j\omega)$ plot has frequency $\omega = \omega_a$ at a. The $G_0/G_c(j\omega)$ plot has $\omega = 1/T_2$ at f. By choosing T_2 properly, it is possible to let the $G_0/G_c(j\omega)$ have the frequency $\omega = \omega_a$ also at a. Using the circle diagram theory, it is done very simply as follows: Draw line ba. Next, draw line be and line ef tangential to the semicircle at b and f, respectively. Now g is the intersection of these two straight lines ba and ef. Then,

$$\frac{\overline{eg}}{\overline{ef}} = \frac{\omega_a}{\frac{1}{T_2}} = \omega_a T_2.$$

Therefore,

$$T_2 = \frac{1}{\omega_a}\frac{\overline{eg}}{\overline{ef}}$$

and

$$T_1 = \overline{od}\,T_2.$$

The design method does not give any answer to the question of the optimum value for T_1/T_2. The decision is left to the designer to make. If he wants to retain the original gain to the highest possible frequency, then the optimum ratio T_1/T_2 is the one that gives a minimum value of T_1. This is very easily found by drawing several circles and finding T_1 for each case.

B. Phase-Lead Compensating Network

In this example, $G_s(j\omega)$ of Fig. 5(c) is the Nyquist plot of the system, and a phase-lead network is used to obtain a specified gain margin. The transfer function of the phase-lead network is

$$G_c(s) = \frac{1 + sT_1}{1 + sT_2} \qquad T_1 > T_2.$$

Actually, this is a combination of an ordinary phase-lead network and an amplifier. The purpose is, of course, to have unity gain at dc.

Fig. 5(b) shows the $G_c(j\omega)$ plot and the $1/G_c(j\omega)$ plot which are again semicircles. The $G_0/G_c(j\omega)$ plot is shown in Fig. 5(c). This semicircle has to intersect the $G_s(j\omega)$ plot in order to have a possible design. The ratio of the two time constants is given by

$$\frac{T_1}{T_2} = \frac{\overline{ob}}{\overline{od}}.$$

Let us find a phase-lead network with a minimum possible ratio of T_1/T_2. The purpose is to keep the gain increase at high frequencies as small as possible. This is very simply done by drawing the $G_0/G_c(j\omega)$ semicircle so that it makes a tangential contact with the $G_s(j\omega)$ plot. Let a be the point where the two plots meet tangentially. Straight lines efg and bag are drawn as in the preceding example. Then,

$$\frac{\overline{eg}}{\overline{ef}} = \frac{\omega_a}{\frac{1}{T_1}} = \omega_a T_1.$$

Therefore,

$$T_1 = \frac{1}{\omega_a}\frac{\overline{eg}}{\overline{ef}}$$

and

$$T_2 = \frac{\overline{od}}{\overline{ob}}\,T_1.$$

The design of the compensating network is is now completed.

When the $G_c(j\omega)$ plot is other than a circle, the design procedure is not as simple as in the above two examples. The new method of compensating network design, however, is still applicable and useful.

Hiroshi Amemiya
Electronic Data Processing Div.
RCA
Camden, N. J.

Perturbation Approach to the Response of a Control System*

A general approach to designing an automatic control system is to design compensation for an assumed plant configuration to yield a desired system response. The choice of plant parameters may be based on their value during a large percentage of the operating period, but in most practical applications, these basic parameters would be expected to shift slightly during the operation of the system. The question then arises as to what has happened to the system response with the values of the parameters varied slightly from those used in the synthesis procedure.

A rather simple approach to the question of what has happened to the system response with small shifts of the basic plant parameters is provided by application of perturbation techniques to the closed-loop transfer function of the system. The basis of this technique is simply the total differential; that is, the total variation of a function is the sum of the partial derivatives of that function with respect to each of its variables, multiplied by the change in each variable, respectively.

* Received by the PGAC, January 9, 1961.

$$\Delta G(a, b, c,) = \frac{\partial G}{\partial a}\Delta a + \frac{\partial G}{\partial b}\Delta b + \frac{\partial G}{\partial c}\Delta c, \quad (1)$$

where Δa, Δb, Δc are small.

Since the Laplace transformation is a linear transformation, it is permissible, and in most cases much simpler, to perform the analysis in the s plane.

To demonstrate the technique, the perturbation analysis would proceed as follows (see Fig. 1):

$$C(s) = R(s)[G_c(s) + \Delta G_c(s)], \quad (2)$$

where

$C(s)$ = the output of the system
$R(s)$ = the input of the system
$G_c(s)$ = the closed-loop transfer function of the system with unperturbed plant parameters
$\Delta G_c(s)$ = the variation of $G_c(s)$ due to small changes in the basic plant parameters

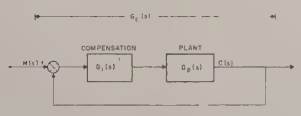


Fig. 1—Block diagram of control system.

$$G_c(s) = \frac{G_1(s)G_p(s)}{1 + G_1(s)G_p(s)}, \quad (3)$$

where

$G_1(s)$ = the compensation, which is assumed constant throughout the operation of the system
$G_p(s)$ = the plant which is a function of several basic parameters, *i.e.*,

$$G_p(s) = G_p(a_1, a_2, a_3, \cdots). \quad (4)$$

Then, taking the total differential of $G_c(s)$, that is, applying (1) to (3) and (4)

$$\Delta G_c(s) = \frac{\partial G_c(s)}{\partial G_p(s)} \sum_i \frac{\partial G_p(s)}{\partial a_i}\Delta a_i, \quad (5)$$

where

$$\frac{\partial G_c(s)}{\partial G_p(s)} = \frac{G_c(s)}{G_p(s)}\left[\frac{1}{1 + G_1(s)G_p(s)}\right]. \quad (6)$$

Then, defining a perturbation transfer function $G_{\Delta a_i}(s)$ as:

$$G_{\Delta a_i}(s) = \frac{1}{G_p(s)[1 + G_1(s)G_p(s)]}\frac{\partial G_p(s)}{\partial a_i}, \quad (7)$$

and substituting (5)–(7) into (2), the system response is

$$C(s) = C_u(s)\left[1 + \sum_i G_{\Delta a_i}\Delta a_i\right], \quad (8)$$

where $C_u(s)$ is the unperturbed system response, *i.e.*,

$$C_u(s) = R(s)G_c(s). \quad (9)$$

Thus, the total system response may be treated as the sum of the unperturbed response and the perturbation response to each basic plant parameter perturbation.

As a first approach to checking the effects of the perturbations on the system response, the maximum variation of the response is simply the evaluation of (8) using the maximum expected perturbations of the parameters. However, a much more sophisticated use of the approach would be an analog computer simulation of the system as depicted in Fig. 2, which would yield the unperturbed system response $C_u(s)$, the perturbation response $C_p(s)$, and the total perturbed response $C(s)$, simultaneously, for any desired variation of the plant parameter. This would allow complete freedom in specifying the environment and consequently evaluating the response of the system to a varying command input while the plant parameters are varying. For example, Δa could be a function representing the change in resistance in the windings of motor due to temperature variation. It could be a continuously varying function or a constant representing the maximum value of Δa.

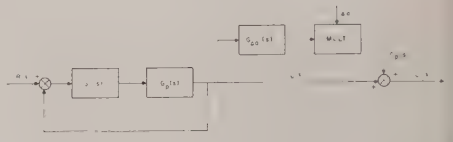


Fig. 2—Block diagram of perturbed control system.

The approach presented here is just one more tool for the control system designer. Obviously, the total perturbed response of the system could be obtained directly by changing the values in the plant itself and then computing the response. However, the method suggested not only yields the total perturbed response of the system, it also isolates the effect of the perturbation of each parameter directly. Thus a much clearer picture of what is actually happening is available. Hopefully, this will lead to a better understanding of the problems involved, and, by displaying the sensitivity of the control system to each parameter, may lead to an improvement in the plant itself by suggesting small changes in its configuration. For example, it could illustrate the necessity of temperature compensating elements.

Paul Mosner
Res. and Advanced Dev. Div.
AVCO Manufacturing Corp.
Wilmington, Mass.

Recent PGAC Chapter Meetings

The following compilation, prepared by Louis B. Wadel, Chairman of the PGAC Chapters Committee, is a list of meetings held by local PGAC chapters during the period January 1, 1959, through those reported in the May, 1961, issue of PROCEEDINGS OF THE IRE.

AKRON

6-16-59—"Adaptive Control Systems," R. N. Bretoi, Minneapolis-Honeywell Regulator Co.

BALTIMORE

2-18-59—"Air Traffic Control Simulator," R. Edwards, Aircraft Armaments, Inc.
3-19-59—"Cobalt 60 Rotational Teletherapy Controls," R. Riley, Westinghouse Electric Corp.
4-29-59—"Multi-Dimensional Servos," N. H. Choksy, The Johns Hopkins University
10-21-59—"Synthesis and Flight Test of a Ballistic Control System," O. Kasfe, The Martin Co.
11-17-59—"Non-Conventional Feedback Control Loop Configurations," J. G. Truxal, Polytechnic Institute of Brooklyn (Brooklyn, N. Y.)
1-10-60—"Adaptability and Ultrastable Systems," N. H. Choksy, The Johns Hopkins University
1-12-60—"An Engineering Presentation of the Second Method of Lyapanov," E. J. Lefferts, The Martin Co.
2- 8-60—"Advancements in Space Vehicle Guidance and Control System Mechanization Techniques," A. O. Buckingham, Westinghouse Electric Corp.
2-16-60—"A New Method of Systems Analysis," R. E. Kalman, RIAS.
3- 9-60—"Sampled Data and Nonlinear Analysis," N. H. Choksy, The Johns Hopkins University
3-22-60—"On the Analysis of Bi-Stable Controls," B. E. Amsler, The Johns Hopkins University
4-21-60—"Random Processes in Automatic Control," G. S. Axelby, Westinghouse Electric Corp.
10- 5-60—"Impressions of Russian Progress in Automatic Controls," Dr. J. M. Mozley, The Johns Hopkins Hospital; Dr. R. E. Kalman, RIAS; and G. Axelby, Westinghouse Electric Corp.
10-10-60—"Application of Linearized Analysis Techniques," W. L. Kinney, Cook Electric Co.
12- 8-60—"Application of Automatic Controls to the Field of Medicine," Dr. S. A. Talbot, The Johns Hopkins Hospital

BOSTON

2- 9-59—"A Discussion of the Problem of Safe Reentry of a Manned Satellite and a Description of the Control System Required to Accomplish This," H. Wexler, AVCO Res. RAD
3- 3-59—"Wide Band Carrier Type Amplification for Electrohydraulic Systems," R. E. Clafin, Jr., Servocontrol Div. of the Oilgear Co.
12- 8-59—"Two-Mode Servo Controlled Weighing System," D. Martini and P. Smith, Feedback Controls, Inc.
"A Unique Target-Position Computer," W. Paradse and D. Stallard, Feedback Controls Inc.
2-24-60—"Automatic Control of Large Antennas with Radio Telescopes," F. J. Mullin, R. S. Wellons, and D. S. Kennedy.
4-18-60—"Analog and Digital Computers in Control Systems," R. B. Wilcox, RCA
2-16-61—"Recoverable Space Probes," M. B. Tragerer, M.I.T. Instrumentation Lab. (Cambridge)

CHICAGO

9- 9-60—"The Automatic Focusing System of the Human Eye," J. Warshawsky, Northwestern University (Evanston, Ill.)
11-11-60—"An Active-Circuit Analogue of a Torque-Reflecting Servosystem," E. J. Miller, Motorola, Inc.

DALLAS—FORT WORTH

1- 8-59—"Digital Control System for Fractionation Tower," G. Post, Genesys Corp.
3-10-59—"Nuclear Reactor Instrumentation," J. R. Gardner, Convair
5- 5-59—"Helicopter Simulator," H. Upton, Bell Helicopter Co.
1-19-60—"Tactics of Air Launching a Ballistic Missile for Satellite Reconnaissance," E. Adams, Convair
3- 7-60—"A Control System for Missile with Swivelable Solid Rocket Motor," L. L. Meyer, Chance Vought Electronics Div.
5- 9-60—"The Molecular Approach to Electronic Computer and Automatic Control Components," R. Biesele, Jr., Shockley Transistor Corp. (Palo Alto, Calif.)
1-23-61—"R.H.-2 Digital Flight Control Computer," R. D. Watson, Bell Helicopter Co. (Hurst, Tex.)
3-14-61—"Ferrites as Circuit Elements for Electrically Controlled Microwave Devices," Dr. W. H. Von Aulock, Bell Telephone Labs., Inc. (Whippany, N. J.)
4-25-61—"Digital Actuator," R. H. Myers, Vought Electronics
5-16-61—"Twin-Gyro Space Vehicle Controller," D. F. Sellers, Vought Electronics

LONG ISLAND

1-20-59—"Hot Gas Servos and Their Application to Missiles," C. Myer, Sperry Gyroscope
3- 3-59—"Network Stabilization of AC Servos," G. Weiss, Polytechnic Institute of Brooklyn
5-19-59—"The Sperry Gyrofin Stabilizer," J. Chadwick, R. Cronemeyer, and D. Price, Sperry Gyroscope
10-15-59—"Automatic Control in the Human Body," P. Suckling, State University of New York
10-20-59—"Electronic Computers in Control Systems," J. Truxal, Polytechnic Institute of Brooklyn
2-23-60—"A Semi Graphical Technique for Designing Third Order Systems," E. Gorczycki, Sperry Gyroscope

LOS ANGELES

1-21-59—"Variational Calculus Principles Applied to Adaptive Flight Control Systems," R. Barron, DODCO
2-13-59—"Statistical Characterization of Control System Nonlinearities," R. B. McGhee, University of Southern California, Hughes Aircraft Co.
3-10-59—"Optimum Synthesis of Multipole Control Systems with Random Processes as Inputs," C. T. Leondes and H.-C. Hsieh, University of California at Los Angeles
4-14-59—"An Analog Computer Method for Automatic Plotting of Root Loci," L. Levine, Hughes Aircraft Co.
5-12-59—"Some Control Problems in Astronautics," R. E. Roberson, Ed., *J. Astronaut. Sci.*, and Assoc. Ed., *Astronaut. Sci. Rev.*
6- 9-59—"Synthesis of Bridged-T Networks Using Root-Locus Techniques," T. A. Savo, Hughes Aircraft Co.
10-14-59—"Attitude Control System for Space Probe Launching Vehicles," H. Low, Space Technology Labs.
"Reaction Wheel Control Systems for Space Vehicles," H. Patapoff and R. W. Froelich, Space Technology Labs.
11-10-59—"Adaptive Flight Control Systems," L. Prince, Minneapolis-Honeywell Regulator Co.

12- 8-59—"Aircraft Flight Control Systems," R. K. Smyth and E. R. Buxton, North American Autonetics

2- 9-60—"Adaptive Automatic Flight Control System," K. C. Kramer, Lear, Inc.

3- 8-60—"Adaptive Control Systems," R. M. Du Plessis, North American Autonetics

4-12-60—"Theory and Practice of Booster Rocket Control," Dr. J. Aseltine, Space Technology Labs.
"The Molecular Approach to Electronic Computer and Automatic Control Components," R. Biesele, Jr., Shockley Transistor Corp. (Palo Alto, Calif.)

5- 6-60—"Adaptive Cross-Correlator," G. W. Anderson and R. Buland, Aeronutronics Corp. (Newport Beach, Calif.)

9-20-60—"Moscow Report—Panel Discussion," Dr. J. A. Aseltine, Aerospace Corp; Dr. A. V. Balakrishnan and A. Rosenbloom, Space Technology Labs.; J. M. Salzer, Ramo-Wooldridge; E. L. Peterson, GE, TEMPO (Santa Barbara, Calif.)

10-11-60—"Introduction to Space Guidance," J. M. Slater, North American Autonetics

11- 8-60—"Adaptive Autopilot Study for a High Performance Interceptor Missile," J. C. Simmons, Douglas Aircraft Co. (Santa Monica, Calif.)

12-13-60— "Guidance and Control Aspects of the Ranger Program," R. Morris, Jet Propulsion Lab. (Pasadena, Calif.)

1-10-61—"Atlas Control System and Mercury Abort and Pilot Safety System," R. Goad and D. R. White, Space Technology Labs.

2-14-61—"Astrodynamics as it is Related to Guidance and Control," R. M. L. Baker, Jr., USAF

MILWAUKEE

10-20-59—"Optimizing the Transient Performance of a Pneumatic Temperature Control System," J. P. Metzger, Milwaukee School of Engrg.

2- 9-60—"Russian Engineering Education with Emphasis on Automatic Control," T. J. Higgins, University of Wisconsin (Madison)

5-10-60—"Feedback Theory Applied to Production and Inventory Control," Dr. G. J. Murphy, Northwestern University (Evanston, Ill.)

11-15-60—"Feedback Controls for Inertially Guided Missiles," R. Brown, AC Spark Plug Div. of General Motors (Oak Creek, Wis.)

2- 7-61—"Automatic Control Applied to Industrial Processes," W. E. Korsan, Allis-Chalmers Co. (West Allis, Wis.)

PHILADELPHIA

10-20-59—"The Compensation of a Digital Type II Servo," R. P. Cheetham, RCA

10-20-60—"The Model Reference Adaptive Control System," H. P. Whitaker, M.I.T. (Cambridge, Mass.)

12- 8-60—"Magnetic Orientation Control of Spin Stabilized Satellites," W. Manger, RCA

Contributors

H. C. Bourne, Jr. (SM'56) was born in Tarboro, N. C., on December 31, 1921. He received the B.S. degree in 1947, the M.S. degree in 1948, and the Sc.D. degree in electrical engineering in 1952, all from the Massachusetts Institute of Technology, Cambridge.

H. C. BOURNE, JR.

He was Assistant Professor of electrical engineering at M.I.T. from 1952 to 1954. He then joined the University of California, Berkeley, as Assistant Professor of electrical engineering, becoming an Associate Professor in 1956.

Dr. Bourne is a member of Tau Beta Pi, Eta Kappa Nu, Sigma Xi, the AIEE, and the American Society for Engineering Education.

❖

F. R. Delfeld was born in Brownsville, Wis., on March 21, 1926. He received the B.E.E. degree from Marquette University, Milwaukee, Wis., the M.S.E.E. degree from the University of Wisconsin, Madison, and the Ph.D. degree from Northwestern University, Evanston, Ill., in 1947, 1950 and, 1960, respectively.

F. R. DELFELD

From 1947 to 1956, he was on the electrical engineering faculty at Marquette University where he held the position of Assistant Instructor. From 1952–1956 he was Assistant Professor of electrical engineering. His activities at Marquette consisted of teaching lecture and laboratory courses in electrical machinery, electronics, communications, and industrial electronics. From 1952 to 1955 he was responsible for the Electronics Laboratory facilities. Concurrent with teaching assignments at Marquette, he was engaged in consulting with the Mosaic Tile Company, Collins Associates, Dittmore-Freemuth Co., Milwaukee County General Hospital and the Veterans Hospital on various technical problems. From 1956 to 1958 he was employed by AC Spark Plug Division of the General Motors Corporation, Milwaukee, as a Project Engineer and Senior Project Engineer in the Systems and Servo Section of the Mace Inertial Guidance Project. His activities included system design and development of the guidance equipment, establishment of performance criteria, and error analysis of the guidance system computer and design studies. From May to September, 1958, he supervised the Systems Design Group of the Mace Inertial Guidance System Project. In 1960, after completing the requirements for the Ph.D. degree, he returned to the AC Spark Plug Division.

Dr. Delfeld is a member of Eta Kappa Nu and Sigma Xi. He is a Registered Professional Engineer in Wisconsin.

❖

Isaac M. Horowitz (S'52–A'53–M'58–SM'60) was born in Safed, Israel, on December 15, 1920. He received the B.S. degree in mathematics and physics from the University of Manitoba, Winnipeg, Can., in 1945; the B.S. degree in electrical engineering from Massachusetts Institute of Technology, Cambridge, in 1952; and the M.S.E.E. and D.E.E. degrees from the Polytechnic

I. M. HOROWITZ

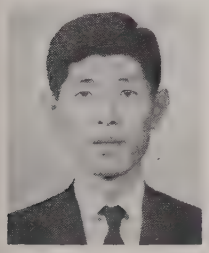
Institute of Brooklyn, Brooklyn, N. Y., in 1953 and 1956, respectively.

From 1956 to 1958, he was an Assistant Professor in the Department of Electrical Engineering at the Polytechnic Institute of Brooklyn. Since 1958, he has been associated with the Hughes Research Laboratories, Malibu, Calif., in the Exploratory Studies Department. He has done research in magnetic amplifiers, active network synthesis, and feedback theory.

In 1956, Dr. Horowitz won the National Electronics Conference award for the best paper presented at the Conference. He is a member of the AIEE.

❖

Takashi Isobe was born in Tokyo, Japan, on January 6, 1914. He received the B.S. degree in physics and the D.Eng. degree from the University of Tokyo in 1937 and 1945, respectively.



T. Isobe

A military technical officer from 1937 through 1945, he joined the faculty of the University of Tokyo as an Instructor in 1945, became an Assistant Professor in 1946, and a Professor of instrumentation in 1948. For the academic year 1959–1960, he was a Visiting Professor at the School of Electrical Engineering, Cornell University, Ithaca, N. Y., on leave of absence from the University of Tokyo. He has contributed many technical papers to Japanese periodicals that are principally concerned with instrumentation and automatic control, and was awarded the Recorder-Controller Section, SAMA Award for his paper, "A New Flowmeter for Pulsating Gas Flow," from the ISA in 1960.

Dr. Isobe is a member of the ISA and Sigma Xi.

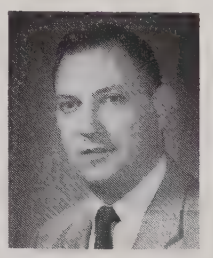
❖

T. T. Kadota was born in Ehime-ken, Japan, on November 14, 1930. He received the B.S. degree in 1953 from Yokohama National University, Yokohama, Japan, and the M.S. degree in 1956 and the Ph.D. degree in 1960 from the University of California, Berkeley, all in electrical engineering. He was an Engineer at Kansai Electric Power Co., Osaka, Japan, from 1953 to 1954, and a Teaching and Research Assistant at the University of California during 1955–1960. He joined Bell Telephone Laboratories, Inc., Whippany, N. J., in 1960.

T. T. Kadota

Dr. Kadota is a member of Sigma Xi.

Robert Kramer (M'55) was born in Quincy, Mass., on April 25, 1927. He received the S.B., S.M., and Sc.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1949, 1952, and 1959, respectively.



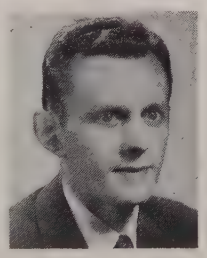
R. Kramer

Since 1949 he has been employed at the Electronic Systems Laboratory (formerly the Servomechanisms Laboratory) of M.I.T. His work has included magnetic particle clutch development, automatic fire control system design and testing, and a broad area of work in the field of torpedo control systems. More recently, the development of thin film devices has been his principal activity. He is presently a Project Engineer and a Lecturer in Electrical Engineering at M.I.T.

Dr. Kramer is a member of Sigma Xi.

❖

Gordon J. Murphy (M'55) was born in Milwaukee, Wis., on February 16, 1927. He received the B.S. degree in electrical engineering from the Milwaukee School of Engineering in 1949. He continued his studies in an evening graduate program at the University of Wisconsin, Milwaukee, and received the M.S. degree in electrical engineering in 1952. He received the Ph.D. degree in 1956 from the University of Minnesota, Minneapolis.



G. J. Murphy

From 1949 to 1951 he was Assistant Professor of electrical engineering at the Milwaukee School of Engineering. In 1951 he accepted a position as Project Engineer on inertial guidance systems at the AC Spark Plug Division of the General Motors Corporation in Milwaukee. In 1952 he accepted a position as Instructor in electrical engineering at the University of Minnesota. There he developed and taught courses in automatic control while continuing his studies as a part-time student. He was appointed Assistant Professor of electrical engineering at the University of Minnesota in 1956. In 1957 he became an Associate Professor of electrical engineering in the Technological Institute of Northwestern University, Evanston, Ill. He has been engaged since then in teaching and in research in the fields of statistical control theory, sampled-data theory, and adaptive control. Since 1960 he has been Professor and Chairman of the Department of Electrical Engineering of Northwestern University.

He is a member of AIEE, ASEE, Sigma Xi, and Eta Kappa Nu.

Takashi Nakada was born in Japan on March 8, 1908. He received the M.S. degree in mechanical engineering in 1932 and the D.Eng. degree in 1944, both from the Tokyo Institute of Technology, Tokyo, Japan.

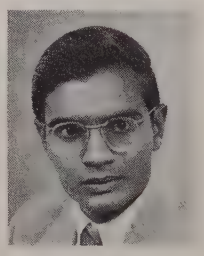
T. Nakada

In 1932 he became Assistant in Mechanical Engineering at the Tokyo Institute of Technology. He subsequently became Assistant Professor in the Research Laboratory of Precision Machinery in 1939. In 1944 he became a Professor of the Tokyo Institute. In 1958–1959 he was Fulbright Visiting Professor at the School of Mechanical Engineering, Purdue University, Lafayette, Ind. In 1960 he was Vice President of the Japan Society of Mechanical Engineers. He returned to the Tokyo Institute of Technology in 1961, where he is currently Director of the Research Laboratory of Precision Machinery and Electronics.

In 1953 Professor Nakada was awarded the Japan Academy Prize for Research on Gears. In 1959 he received a prize from the Japan Society of Mechanical Engineers for his paper on "Feedback Control Increases the Accuracy of Machine Tools."

❖

Kumpati S. Narendra (M'60) was born in Madras, India, on April 14, 1933. He received the B.E. degree from Madras University in 1954, and the M.S. and Ph.D. degrees in applied physics from Harvard University, Cambridge, Mass., in 1955 and 1959, respectively.



K. S. Narendra

From 1959 to 1960 he was a Research Fellow at Harvard, and in 1961 he became a Lecturer in automatic control theory; in the same year he also became an Assistant Professor of applied physics at Harvard. He is a Consultant for the Minneapolis-Honeywell Regulator Company, Boston Division, Mass.

Dr. Narendra is a member of Sigma Xi and the AIEE.

❖

Rufus Oldenburger was born in Grand Rapids, Mich., on July 6, 1908. He received the B.A. degree in Latin and Greek in 1928, and the M.S. and Ph.D. degrees in mathematics in 1930 and 1934, respectively, all from the University of Chicago, Ill.

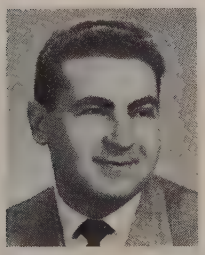
He served with the Woodward Governor Company from 1942–1957, when he left his post as Director of Research to accept a Professorship of electrical and mechanical

engineering at Purdue University, Lafayette, Ind. He is now Professor of mechanical engineering at Purdue, and an Industrial Consultant. He has held professorial chairs and other posts in mathematics from 1930–1950 at the University of Michigan, Ann Arbor; Case Institute of Technology, Cleveland, Ohio; Illinois Institute of Technology, Chicago; De Paul University, Chicago; and the Institute for Advanced Study, Princeton, N. J. He has lectured in several foreign languages and was Visiting Professor in universities in Mexico, France, and Japan. He introduced modern scientific techniques to the prime mover governor industry, discovered and developed various areas of linear and nonlinear control theory and applications, and developed hydraulic circuit theory and the theory of convergence of iteration methods currently used in computer and statistical fields.

R. OLDENBURGER

❖

Elijah Polak was born in Bialystok, Poland, on August 11, 1931. He received the B.E.E. degree in 1956 from the University of Melbourne, Australia, and the M.S.E.E. degree in 1959 from the University of California, Berkeley.



E. POLAK

In 1957 he was an Instrument Engineer with the Imperial Chemical Industries of Australia and New Zealand in Melbourne, Australia, working on the development of transducers for automatic process control. He held summer employment with the National Cash Register Co., Dayton, Ohio, in 1958, and with the IBM Research Laboratories, San Jose, Calif., in 1959 and 1960. His work involved switching circuits and hydraulic servos. At present he is an Associate in Electrical Engineering at the University of California, Berkeley, where he has taught various courses in electrical engineering, and where he is completing work towards the Ph.D. degree in electrical engineering.

Mr. Polak is a member of Sigma Xi.

❖

Charles D. Pollak was born in Detroit, Mich., on October 11, 1930. He received the B.S. degree in 1952 from the U. S. Naval Academy, Annapolis, Md., and the M.S.E.E. degree in 1960 from the U. S. Naval Postgraduate School, Monterey, Calif.

C. D. POLLAK

His professional career in the Navy has been concerned with destroyers and submarines. He currently has the rank of Lieutenant and is designated as qualified for command of submarines. His major field of study at the U. S. Naval Postgraduate School was in the area of Guided Missiles. Subsequently he attended the Polaris Weapons Officers course and is presently assigned as Weapons and Missiles Officer on the nuclear-powered Polaris submarine USS *John Marshall*, under construction in Newport News, Va.

❖

Manoel Sobral, Jr., was born in Salvador, Ba., Brazil, on January 29, 1935. After receiving the B.S. degree from Instituto Tecnologico de Aeronautica, S. Jose dos Campos, S. P., Brazil, in 1958, he spent two years with the Control and Conversion Group of the Department of Electronics of the same Institute.

M. SOBRAL, JR.

He is presently working towards the M.S. degree at the University of Illinois, Urbana, where he is Research Assistant in the Coordinated Science Laboratory.

Kenneth Steiglitz was born in Weehawken N. J., on January 30, 1939. He received the B.E.E. degree, magna cum laude, from New York University, N. Y., in 1959.

K. STEIGLITZ

Upon graduation in 1959 he was associated with the Research Division, College of Engineering, New York University, as an Assistant Research Scientist, engaged in studies of digital-display techniques for radio direction-finding systems. During the academic year 1959–1960, he was a National Science Foundation Fellow at New York University, where he obtained the M.E.E. degree in June, 1960. At present, he is a Teaching Fellow in electrical engineering at New York University, where he is studying for the Sc.D. degree.

Mr. Steiglitz is a member of Eta Kappa Nu and Tau Beta Pi.

❖

George J. Thaler (M'60) was born in Baltimore, Md., on March 15, 1918. He received the B.E. degree and the Dr.Eng. degree from the Johns Hopkins University, Baltimore, Md., in 1940 and 1947, respectively.

G. J. THALER

From 1941 to 1947, he was Instructor and Research Assistant at the Johns Hopkins University. From 1947 to 1951, he was Assistant Professor of Electrical Engineering at the University of Notre Dame, South Bend, Ind. Since 1951 he has been on the faculty of the U. S. Naval Postgraduate School, Monterey, Calif., where he is now Professor of electrical engineering. His major field of interest is feedback control theory, and he has written or co-written four books and over twenty research papers in that area.

Dr. Thaler is a member of Sigma Xi, ASEE, and AIEE. He is a Licensed Professional Engineer.

Announcements

BINDERS FOR THE TRANSACTIONS

Binders of spanish-grain maroon fabrikoid with gold lettering, ruggedly built, with mechanism for holding copies of the TRANSACTIONS in place, are available (4-inch capacity, 24 steel blades). Copies are not damaged by their insertion; each individual copy will lie flat when the pages are turned, and the copies can be removed from the binder in a few seconds. These sturdy binders will protect your file of IRE publications from damage and loss.

$3.00—Standard Binder.....
$3.50—Your Name Imprinted.....
(Indicate exact imprinting of name)
$3.50—Group Name Imprinted.....
(Indicate exact imprinting of group name)
$3.50—Year Imprinted.....
(Specify year)
$4.00—Year and Your Name Imprinted.....
(Specify year)
.....
(Indicate exact imprinting of name)
$4.00—Year and Group Name Imprinted.....
(Specify year)
.....
(Indicate exact imprinting of group name)
$4.50—Year, Your Name and Group Name Imprinted.....
(Specify year)
.....
(Indicate exact imprinting of name)
.....
(Indicate exact imprinting of group name)

Binders may be ordered from:

The Institute of Radio Engineers, Inc.
1 East 79 Street, New York 21, N. Y.

CALL FOR PAPERS

1962 JOINT AUTOMATIC CONTROL CONFERENCE

The 1962 JACC will be held at New York University, University Heights, New York City, on June 27–29, 1962, and will be jointly sponsored by the IRE-PGAC, AIEE, ISA, AIChE, and ASME. Innovations in paper review and documentation procedures are expected to result in a program of outstanding quality and documentation.

The host university features an outstanding control engineering faculty, headed by Dean John R. Ragazzini. The conference General Chairman is Dr. Arthur S. Robinson, Director of Research, Kollsman Instrument Corporation, 80-08 45th Avenue, Elmhurst 73, N. Y. Anthony J. Hornfeck, Director of Research, Bailey Meter Company, 1050 Ivanhoe Road, Cleveland 10, Ohio, is Program Chairman. The IRE-PGAC representative on the JACC Steering Committee is Prof. J. H. Mulligan, Electrical Engineering Department, College of Engineering, New York University, New York, N. Y.

At the 1961 Boulder, Colorado Joint Automatic Control Conference, the 1962 Program Committee set the following paper deadlines for the 1962 Conference:

Abstracts by October 15, 1961
Paper text by November 15, 1961.

Authors will be notified of paper disposition by February 15, 1962 and will have until April 15, 1962, to submit revised manuscripts. Prospective authors of significant papers covering control theory, applications or components are invited to submit abstracts and text through their sponsoring society. Early submission is encouraged.

Abstracts (4 copies) and complete papers (4 copies) from IRE authors should be submitted to the IRE-PGAC Program Committee Representative:

Harold Levenstein
Servo Corporation of America
111 New South Road
Hicksville, L. I., N. Y.

SECOND INTERNATIONAL CONGRESS ON INFORMATION PROCESSING

The Second International Congress on Information Processing will be held in Munich, Germany, from August 27 to September 1, 1962. This Congress is being organized by IFIPS—the International Federation of Information Processing Societies—of which the American representation is the National Joint Computer Committee, NJCC, of the AIEE, IRE, and ACM.

Additional information such as pictures taken at the organizing meeting in Darmstadt in February, 1961, biographical information or pictures of the American arrangements personnel, and the makeup of the American Arrangements Committee will be available soon.

Supplementary material can be obtained either from I. L. Auerbach, Auerbach Electronics Corporation, 1634 Arch Street, Philadelphia 3, Pa., *President* of IFIPS, or from E. L. Harder, Westinghouse Electric Corporation, East Pittsburgh, Pa., *Chairman*, American Arrangements for the Congress.

Call for Papers

The Congress will cover all aspects of Information Processing and Digital Computers including the following:

1) *Business Information Processing*
e.g., data processing in commerce, industry, and administration.
2) *Scientific Information Processing*
e.g., numerical analysis; calculations in applied mathematics, statistics, and engineering; data reduction; problems in operations research.
3) *Real Time Information Processing*
e.g., reservation systems; computer control; traffic control; analog-digital conversion.

4) *Storage and Retrieval of Information*
   e.g., memory devices; library catalogs.
5) *Language Translation and Linguistic Analysis*
6) *Digital Communication*
   e.g., encoding; decoding; error detecting and error correcting codes for digital data transmission.
7) *Artificial Perception and Intelligence*
   e.g., pattern recognition; biological models; machine learning, automata theory.
8) *Advanced Computer Techniques*
   e.g., logical design; logical elements, storage devices; ultra-high-speed computers; program techniques; ALGOL.
9) *Education*
   e.g., selection and training of computer specialists; training of nonspecialists in the use of computers; information processing as a University subject.
10) *Miscellaneous Subjects*
   e.g., growth of the information processing field.

In each category it is planned to cover, where appropriate, the applications of digital computers, programming, systems design, logical design, equipment, and components.

Those wishing to offer papers are invited to send abstracts of 500–1000 words to: Dr. E. L. Harder, Westinghous Electric Corporation, East Pittsburgh, Pa., by *September 15, 1961.* These abstracts will be considered by the international program committee of IFIPS, and authors of selected abstracts will be invited to submit their complete papers (in French or English) for consideration by the program committee in March, 1962.

In addition to accepted papers, there will be invited papers, symposia, and panel discussions.

PRELIMINARY CALL FOR PAPERS
SECOND IFAC CONGRESS

The American Automatic Control Council invites authors to submit papers for the Second Congress of the International Federation of Automatic Control and requests that they signify their intention of doing so as soon as possible. The Second Congress of the IFAC will be held in Basel, Switzerland, in September, 1963, on the invitation of the Swiss Association of Automatic Control.

The total number of papers will be limited to one hundred and there will be no national quotas of papers as in the past. The final selection will be by the IFAC Committee and not the national committee submitting the paper.

The majority of the papers on the program will be comprised of material on Theory and Application of Automatic Controls. A few papers will be accepted on Components, Bibliography, Terminology, and Education.

Prospective authors should write directly to the chairman of the appropriate committee. The Committee *Chairman* will provide the authors with details concerning deadlines, length of paper, etc. Each committee has a member from each of the five societies which provide the membership of AACC: IRE, AIEE, ASME, ISA, and AIChE. The IRE is represented on the AACC by the Professional Group on Automatic Control.

Dr. John Truxal
Chairman, Theory Committee
Electrical Engineering Dept.
Brooklyn Polytechnic Institute
333 Jay Street
Brooklyn 1, N. Y.

Prof. Irving Lefkowitz
Chairman, Application Committee
Mechanical Engineering Dept.
Case Institute of Technology
University Circle
Cleveland 6, Ohio

Prof. J. L. Shearer
Chairman, Component Committee
Dept. of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge 39, Mass.

Prof. T. J. Higgins
Chairman, Bibliography Committee
Engineering Dept.
University of Wisconsin
Madison, Wis.

Dr. H. L. Mason
Chairman, Terminology Committee
National Bureau of Standards
Washington 25, D. C.

Prof. J. H. Mulligan
Chairman, Education Committee
Electrical Engrg. Dept.
New York University
New York 53, N. Y.

CORRECTION

In the Introduction to the May, 1961, issue of these TRANSACTIONS, which featured the 1961 JACC papers, D. L. Lippitt's name was inadvertently omitted. Mr. Lippitt was an alternate 1961 JACC program representative and was responsible for part of the paper review and selection. We wish to acknowledge our thanks to him for his efforts.—*The Editor.*